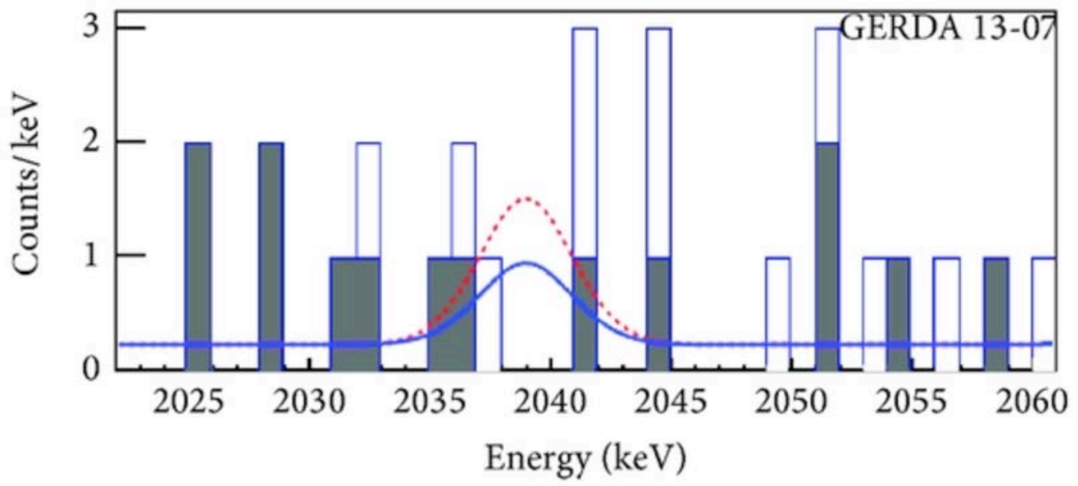
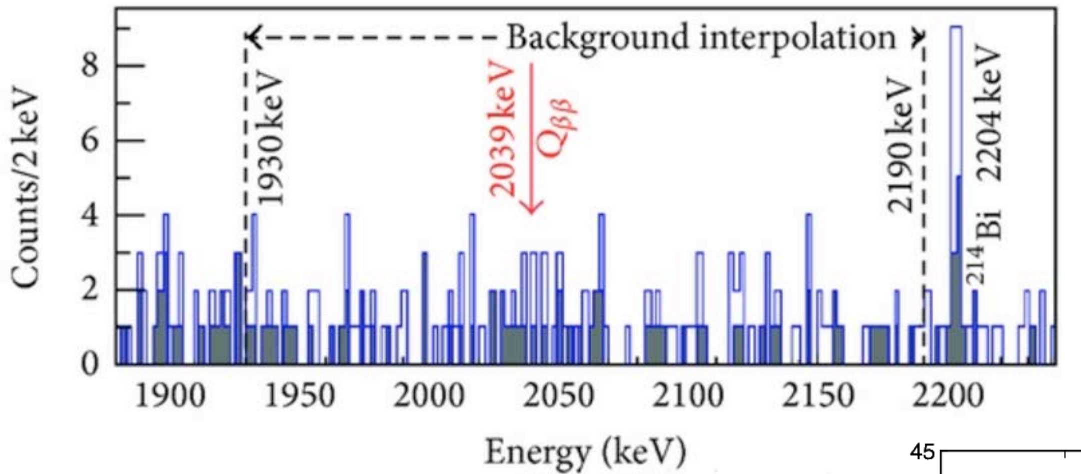
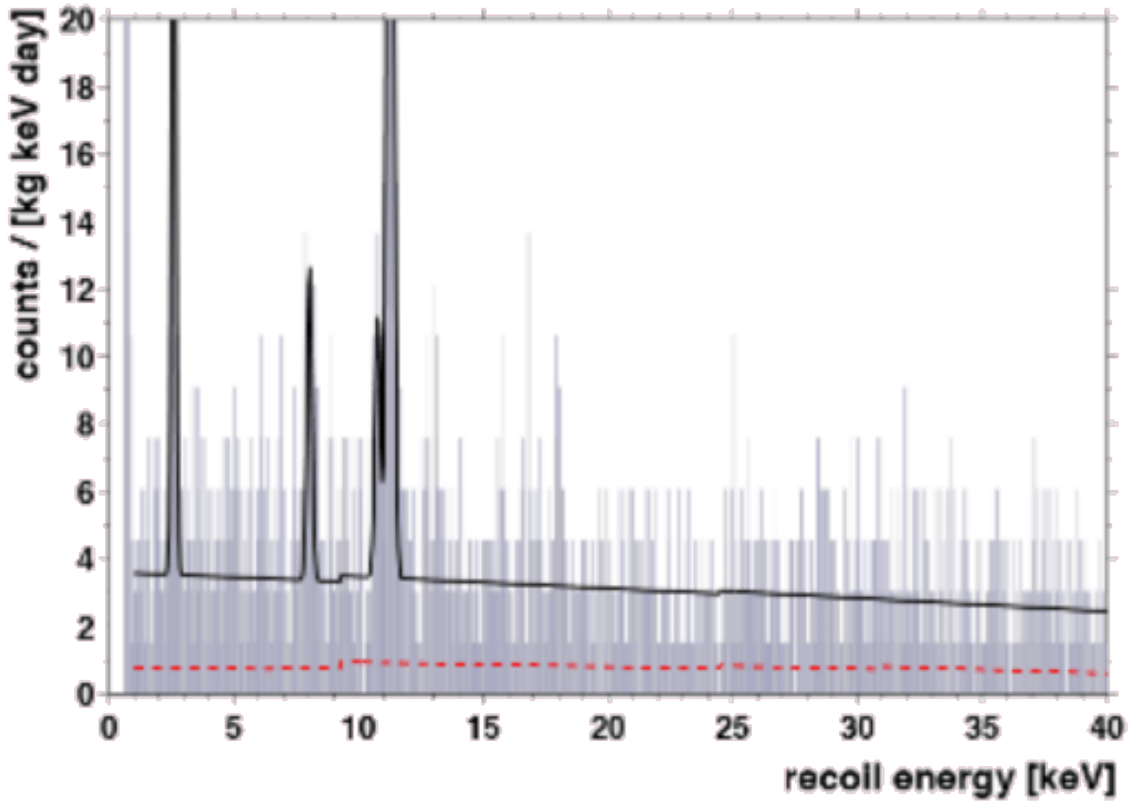
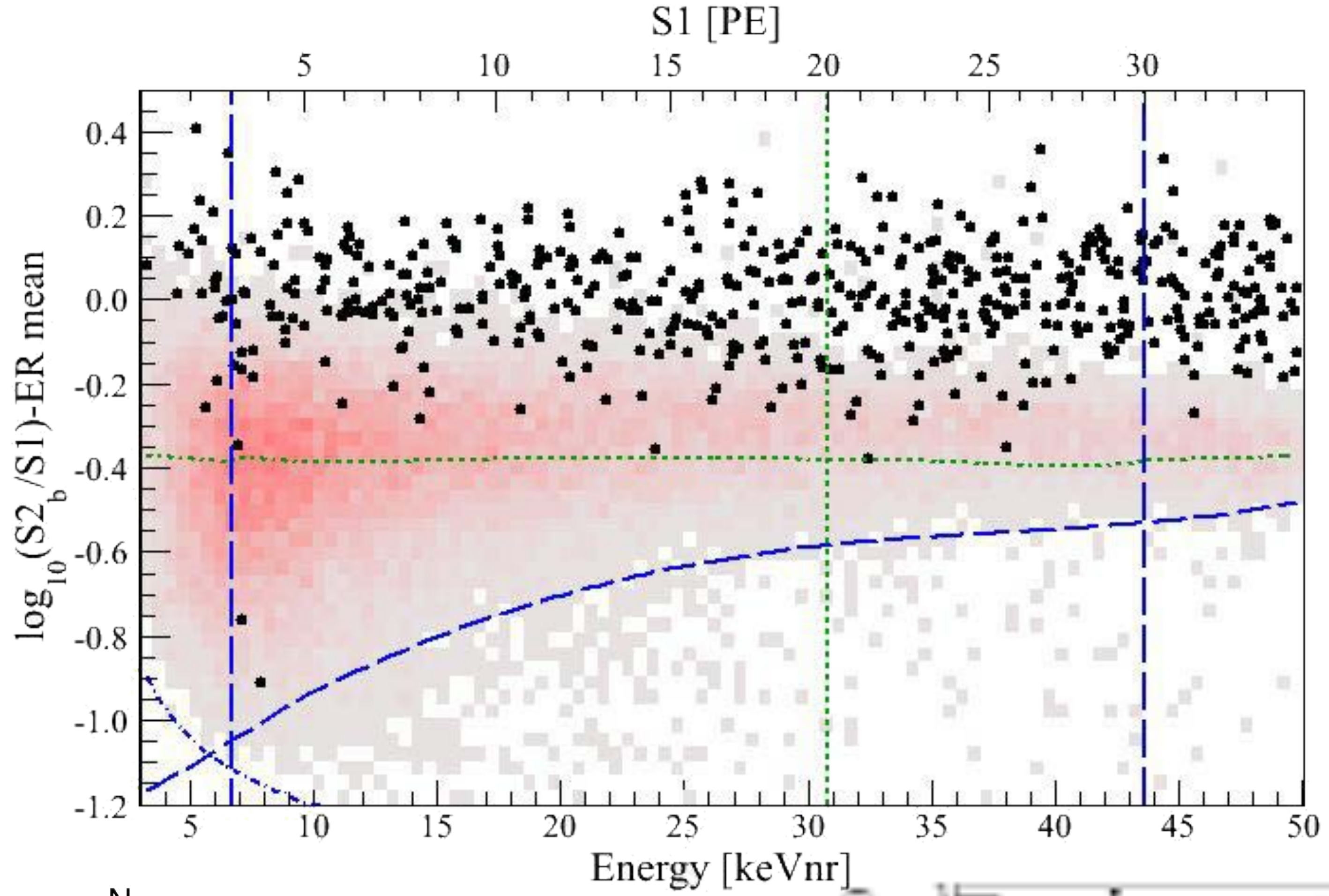


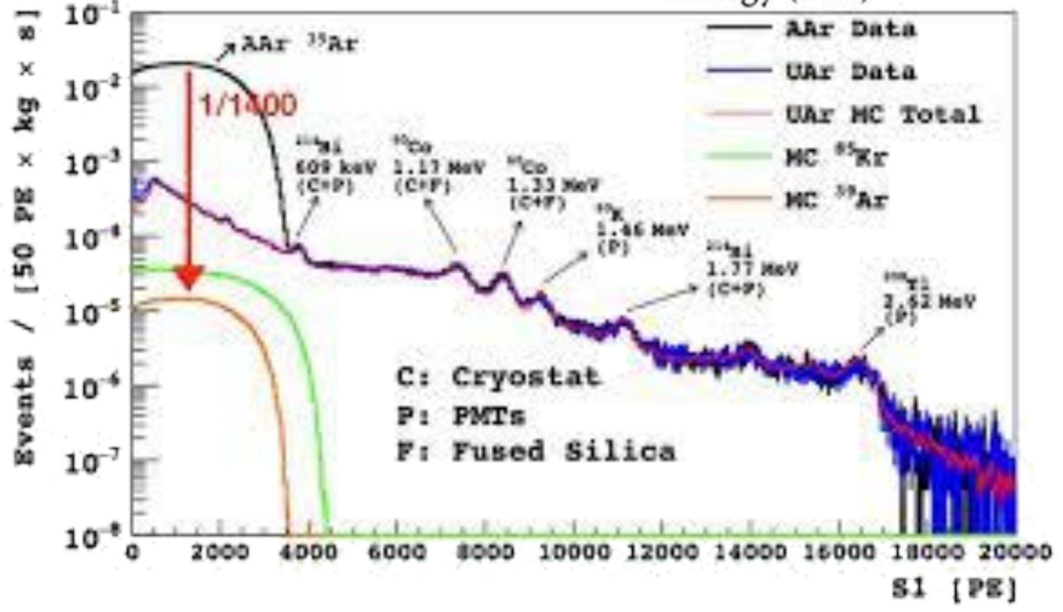


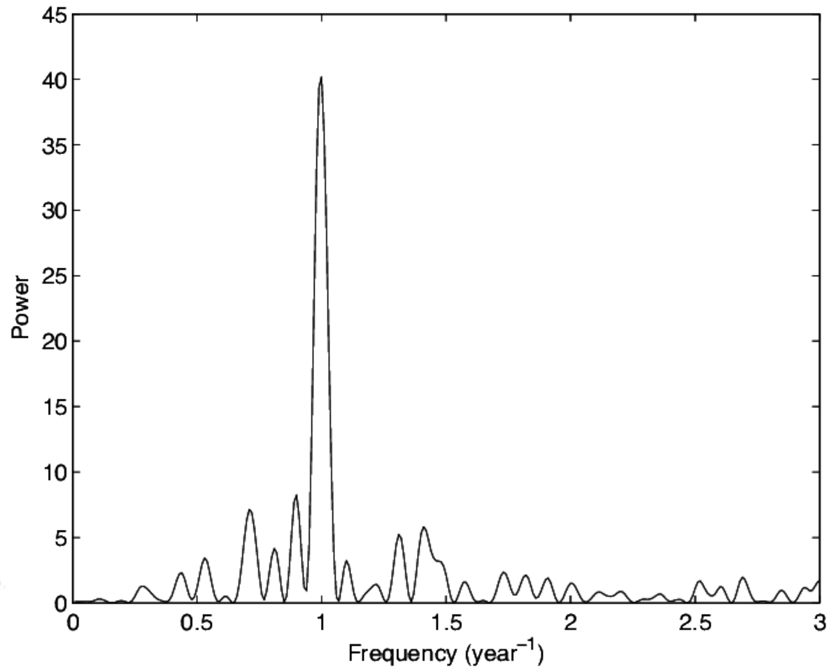


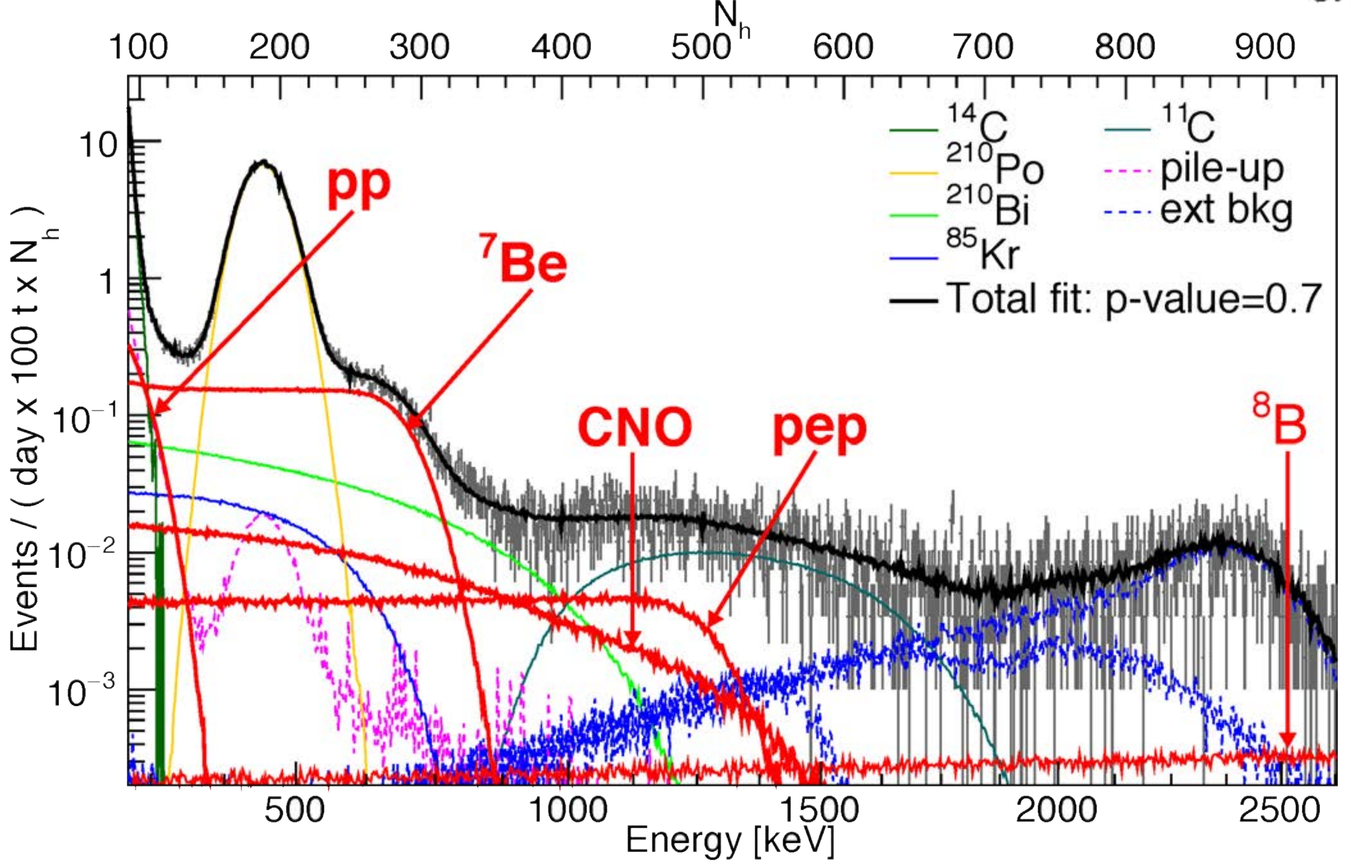


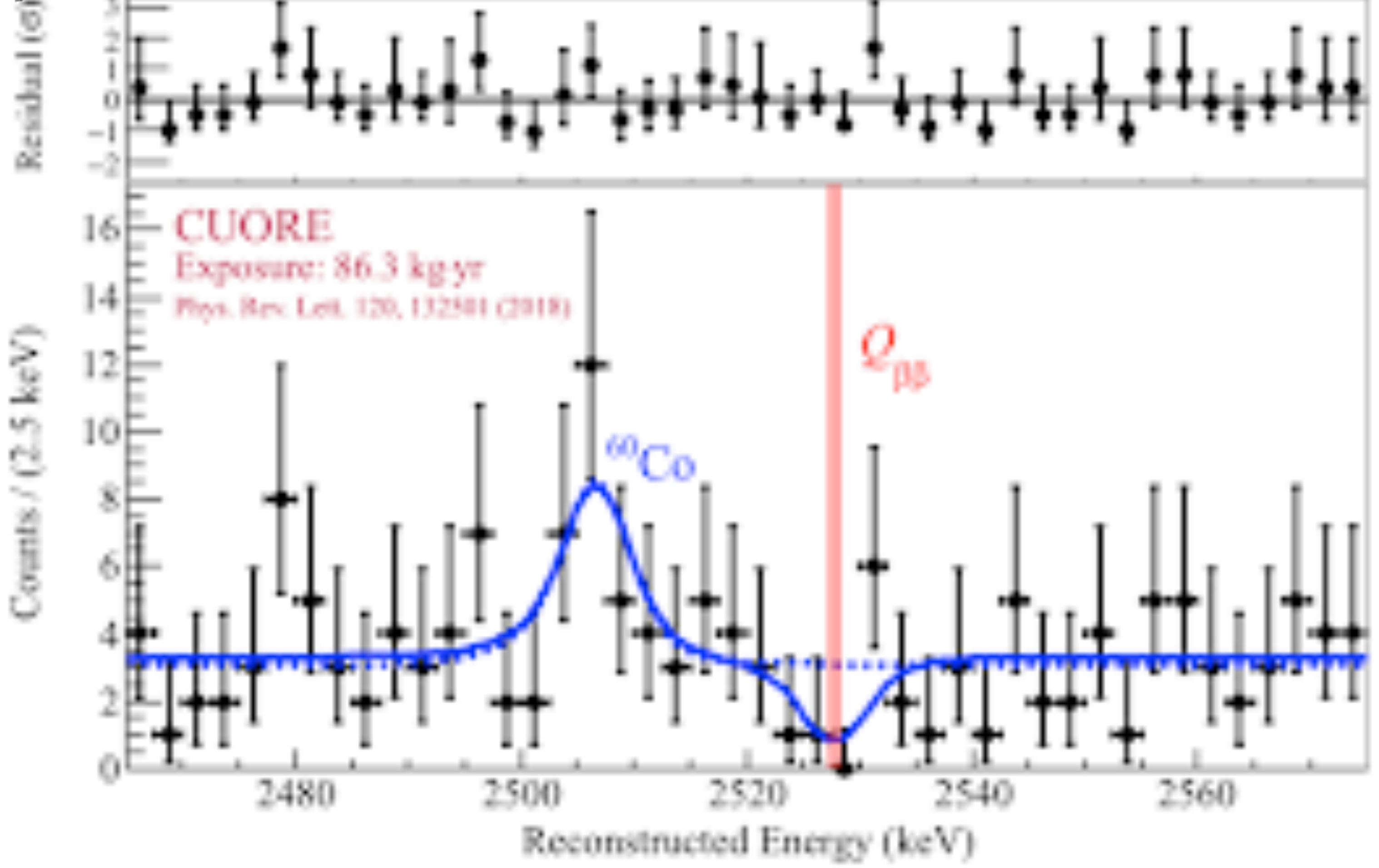
INFN

Istituto Nazionale di Fisica Nucleare









Statistics for Underground Physics

O. Cremonesi - INFN Sez. Milano Bicocca

SOUP2021 - June 28, 2021

Outline

- Fundamentals
  - Measures
  - Probability
  - Random variables
  - Pdfs
  - Bayes
- Likelihood ratio
  - Parameter estimation
  - Results combination
- Statistical tests
  - p-values
  - Limits
- Bayesian methods
  - MCMC, systematic errors

Disclaimer

I am not an expert in statistics

Maybe a practitioner with a longstanding experience in experiments carried out underground

Material drawn from R.Barlow, G.D’Agostini, J.Orear, G.Cowan and F.James

Some books

- R.J. Barlow, Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences, Wiley, 1989
- G.D'Agostini, Bayesian Reasoning in Data Analysis: A Critical Introduction, World Scientific Publishing 2003.
- Luca Lista, Statistical Methods for Data Analysis in Particle Physics, Springer, 2017.
- F. James., Statistical and Computational Methods in Experimental Physics, 2nd ed., World Scientific, 2006
- G. Cowan, Statistical Data Analysis, Clarendon, Oxford, 1998

Statistics

Why we need it

From theory to experiment

Theory (model, hypothesis):

$$\vec{F} = -G\frac{mM}{r^2}\hat{r}$$

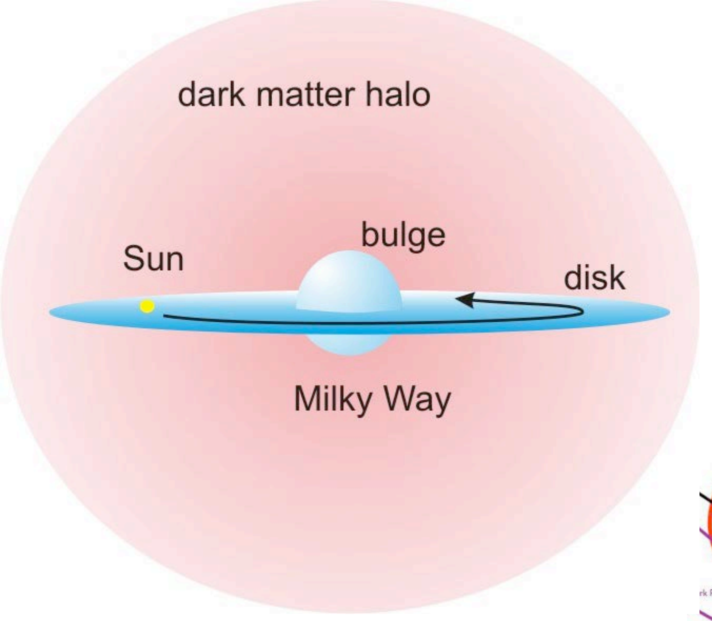
$$F = kx$$

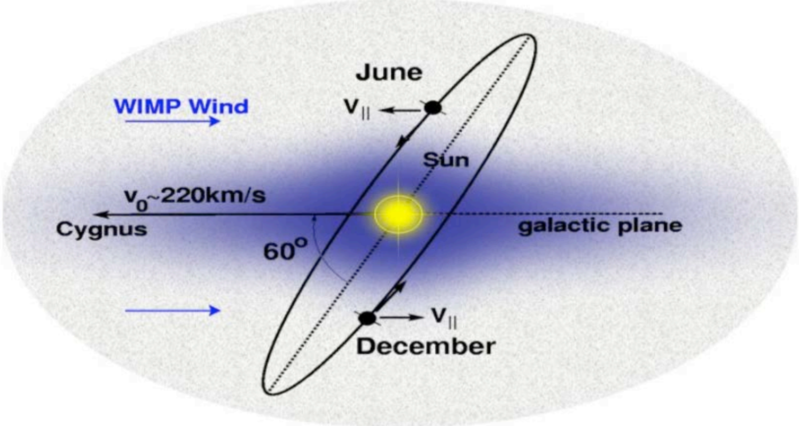
Experiment:

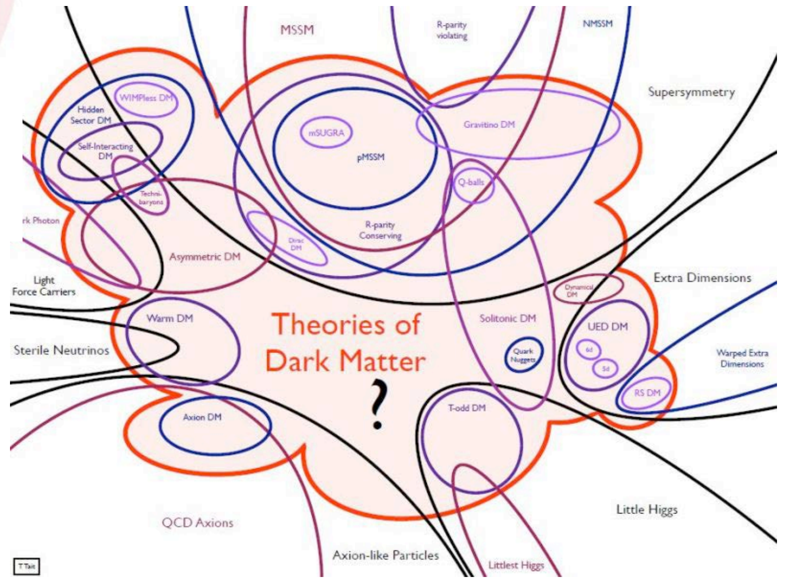
Data selection

From theory to experiment

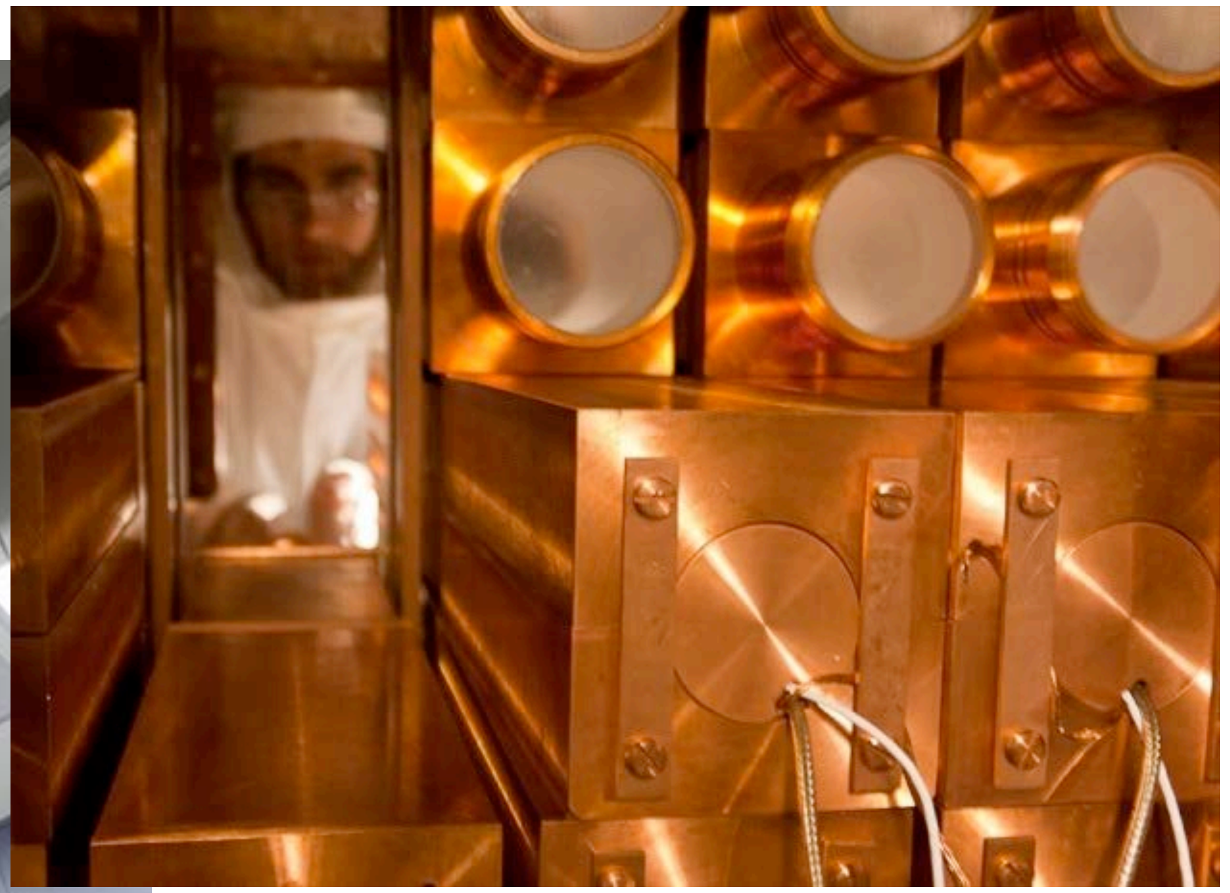
Theory (model, hypothesis):



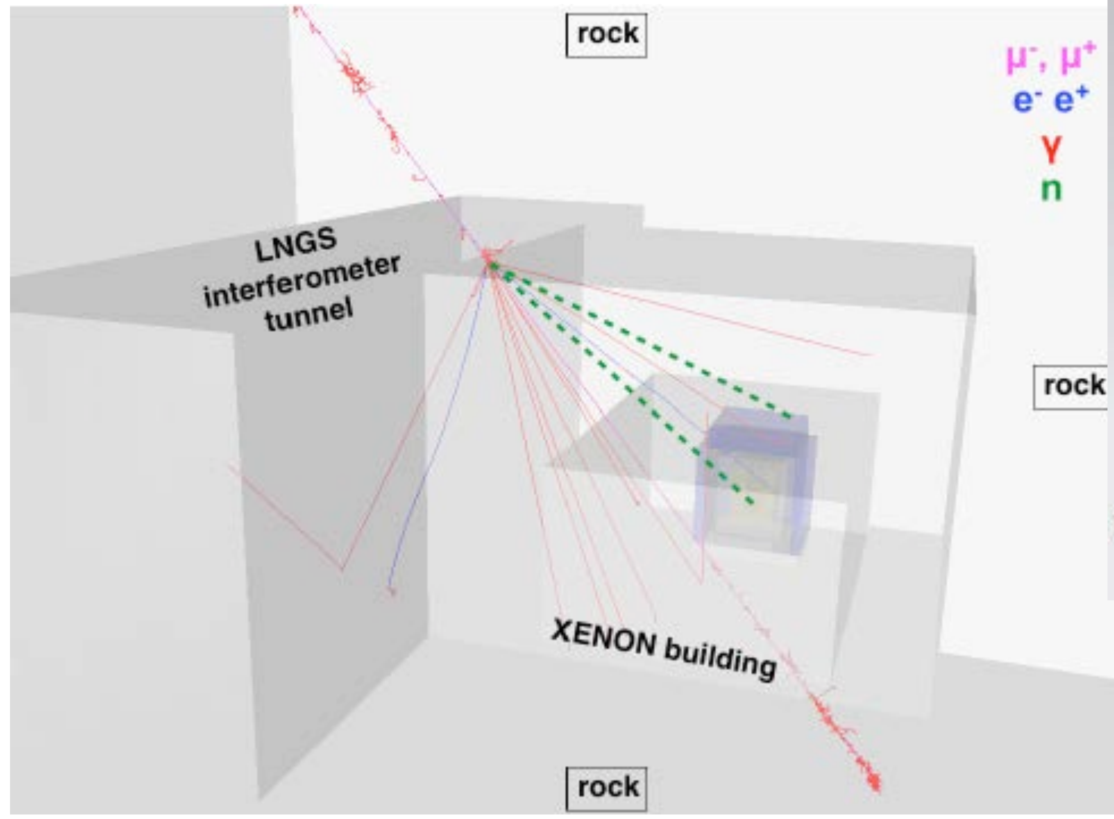


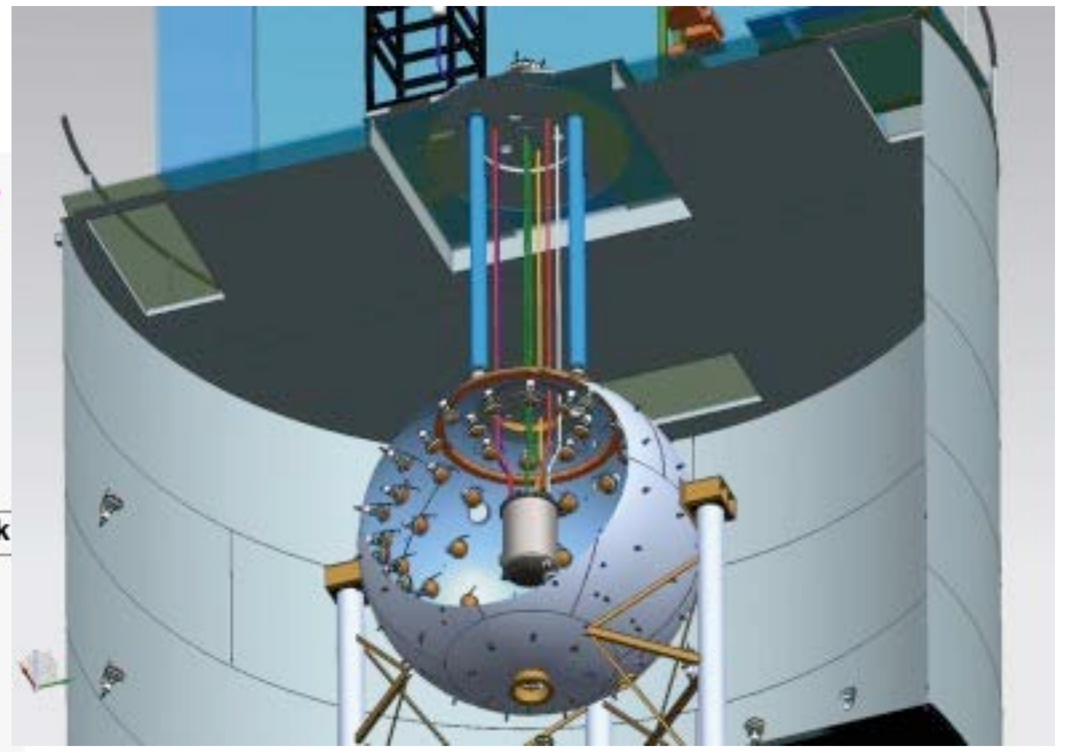


Experiment:

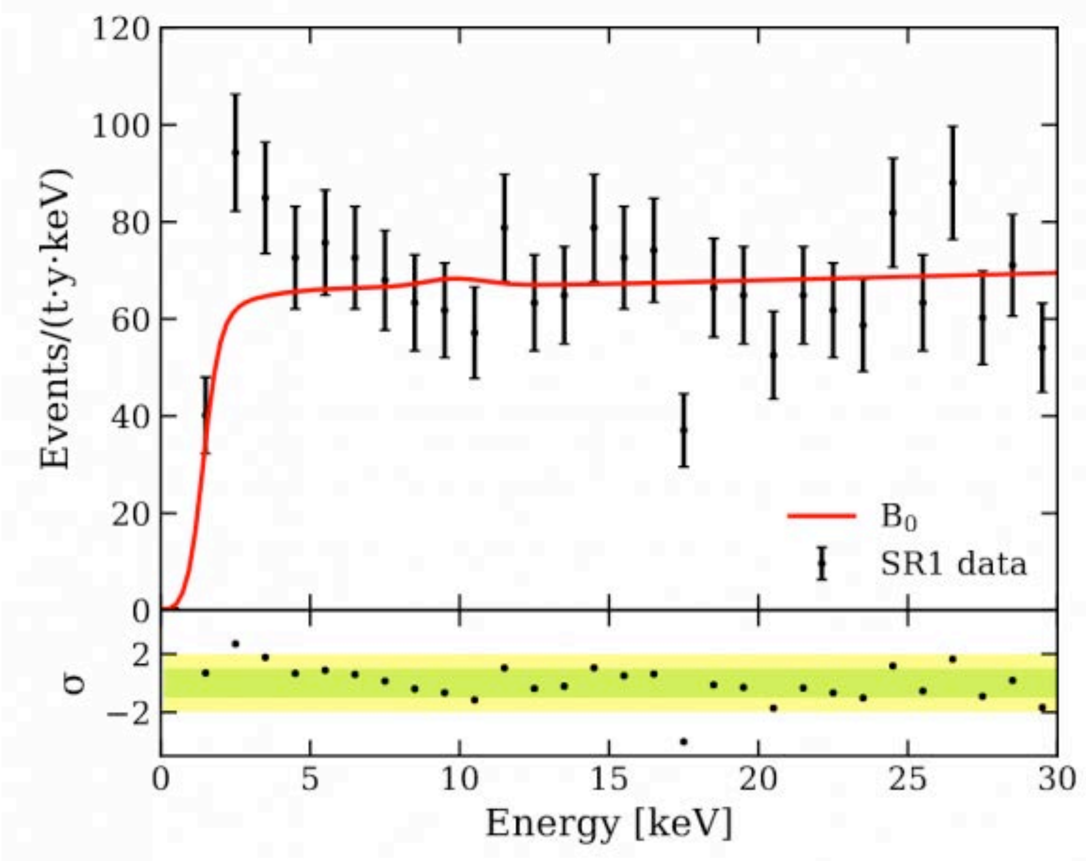


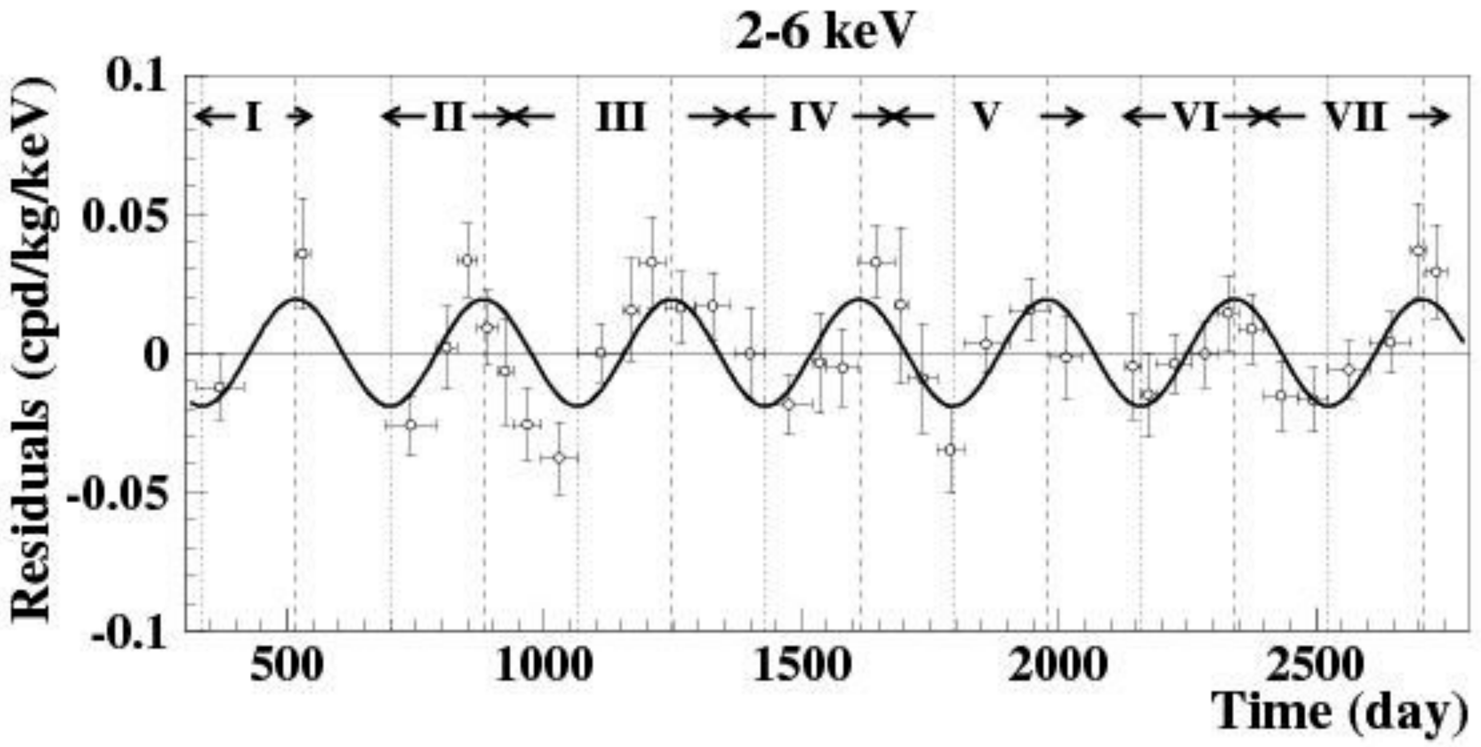
Simulation of detector and cuts





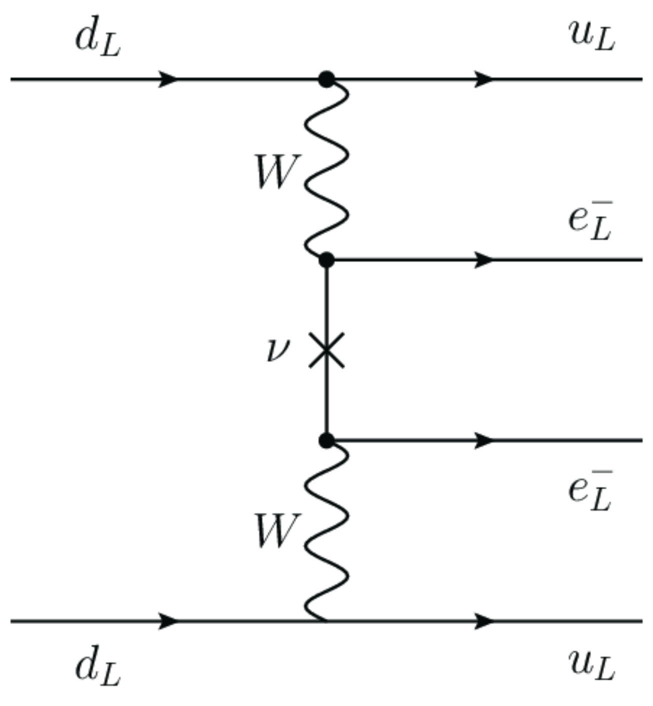
Data selection

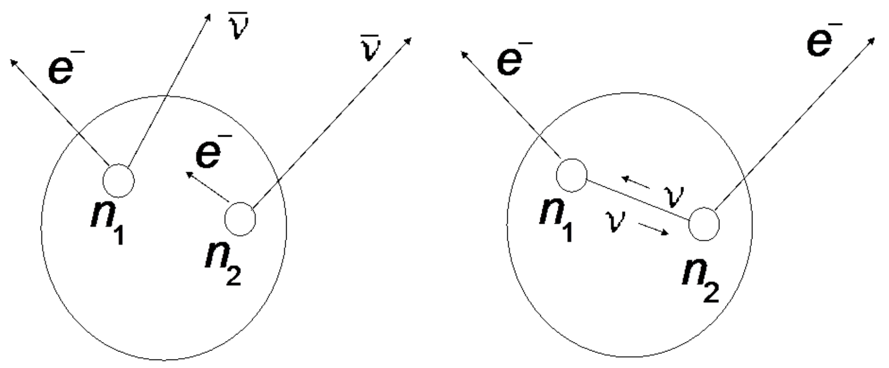


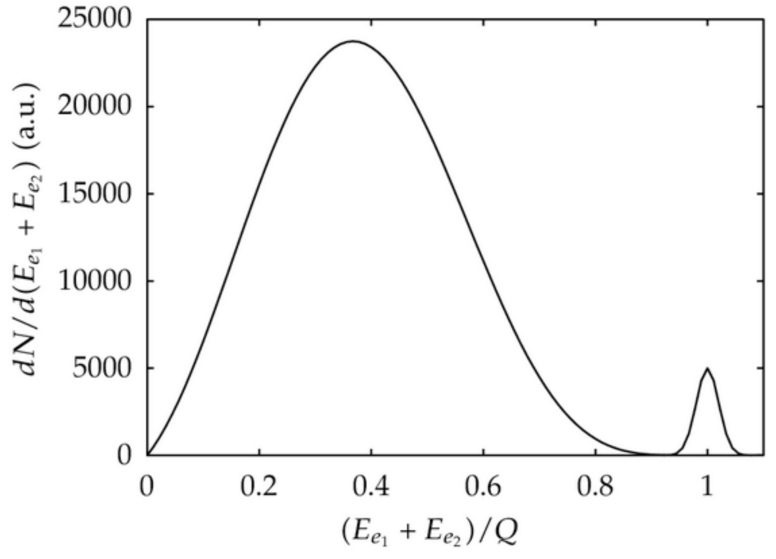


From theory to experiment

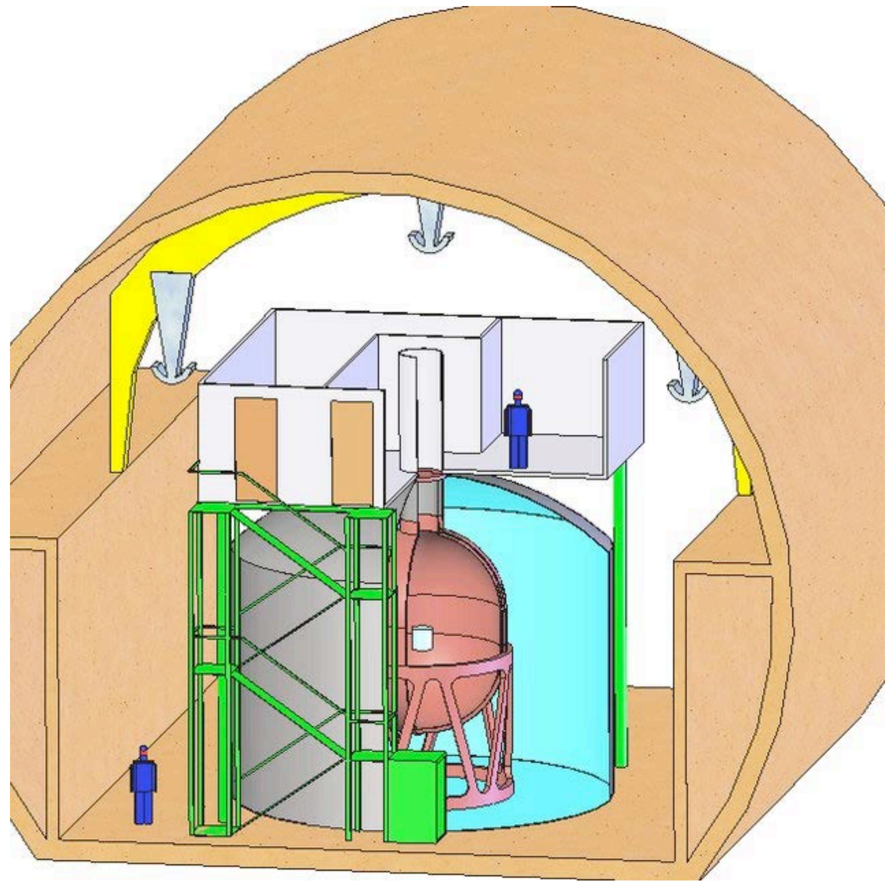
Theory (model, hypothesis):

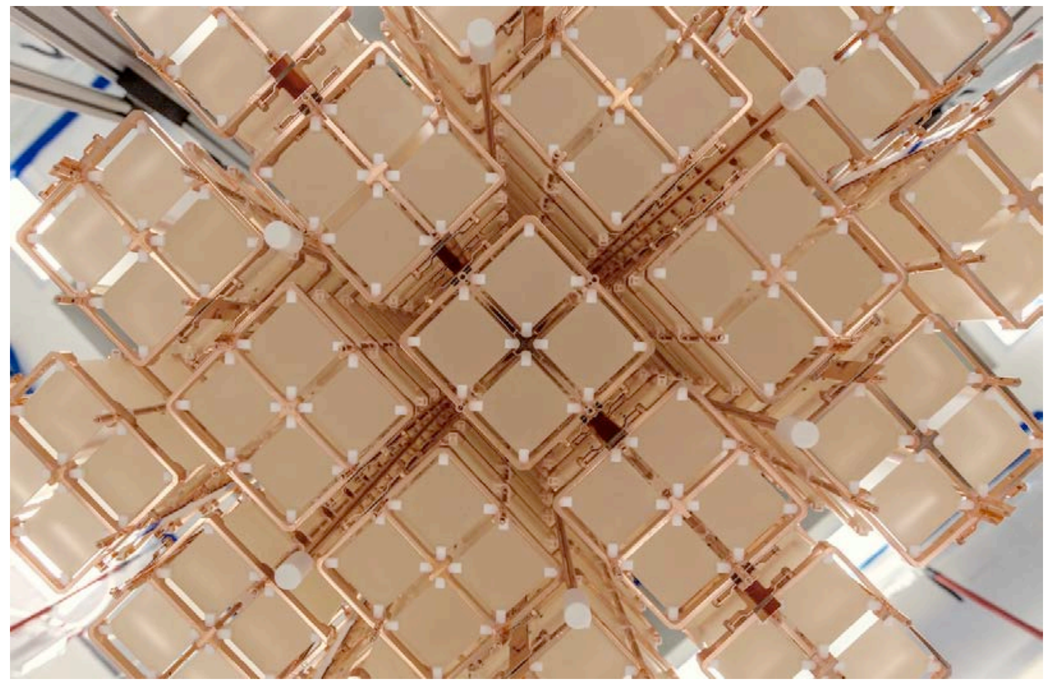




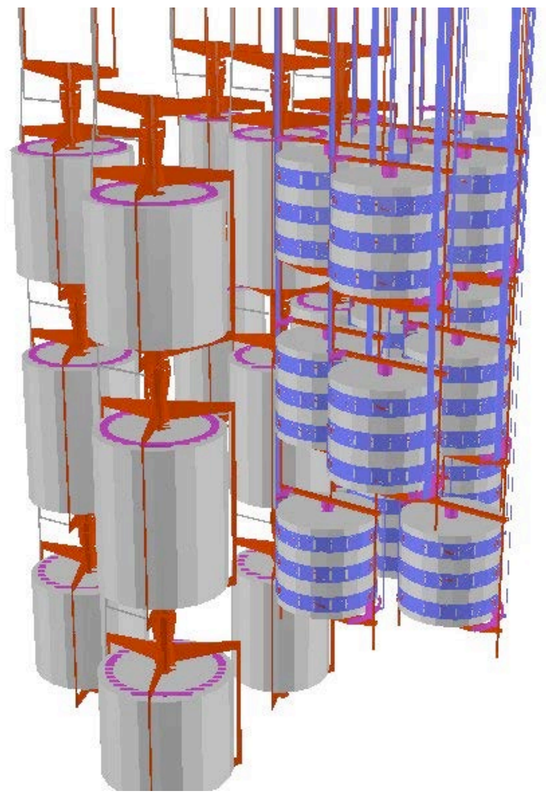


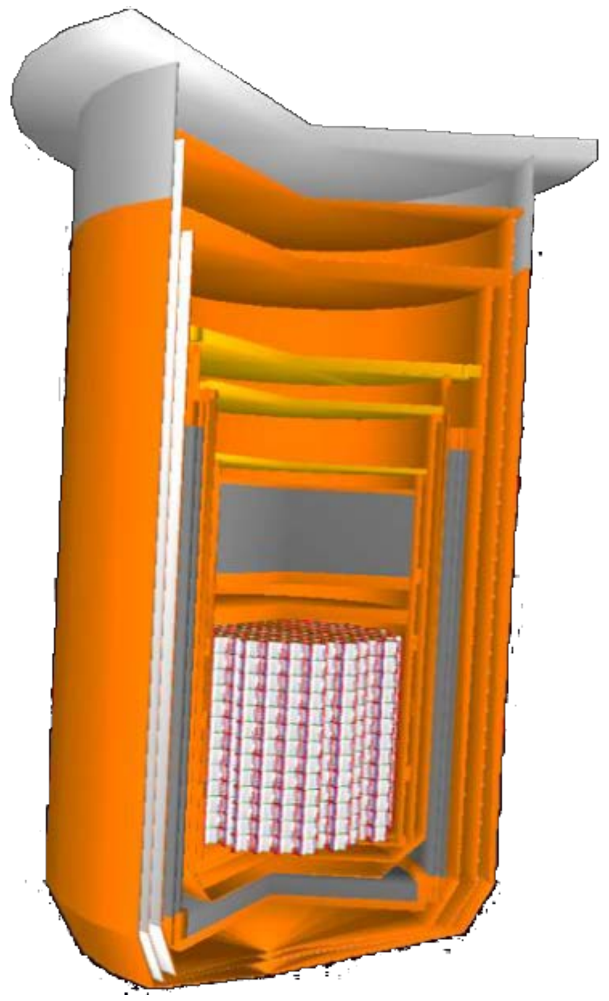
Experiment:



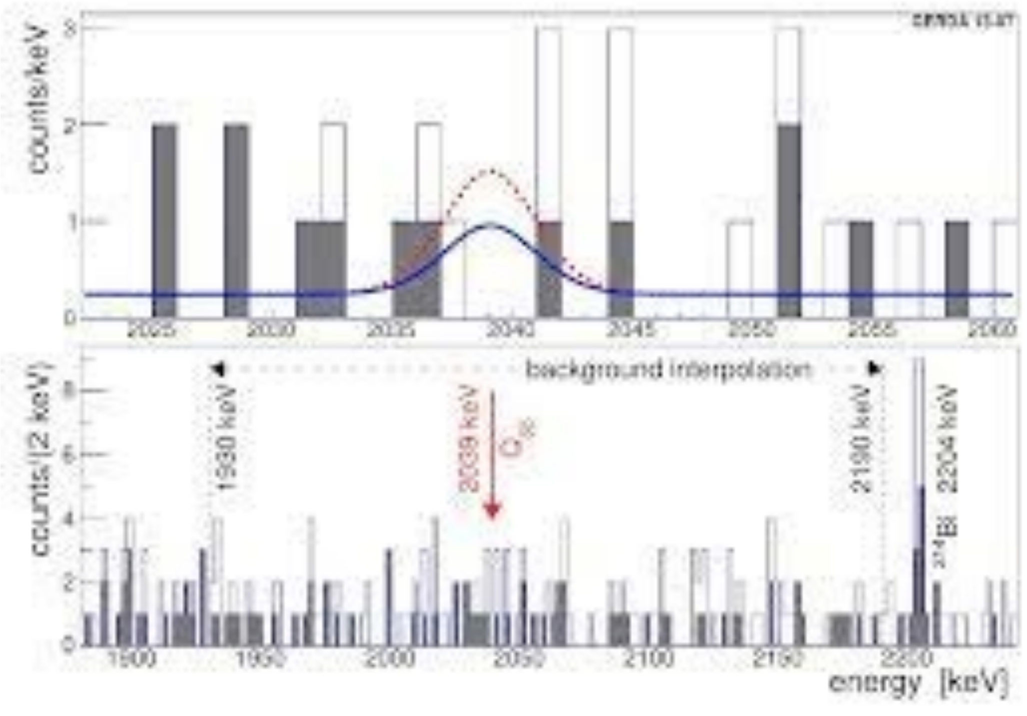


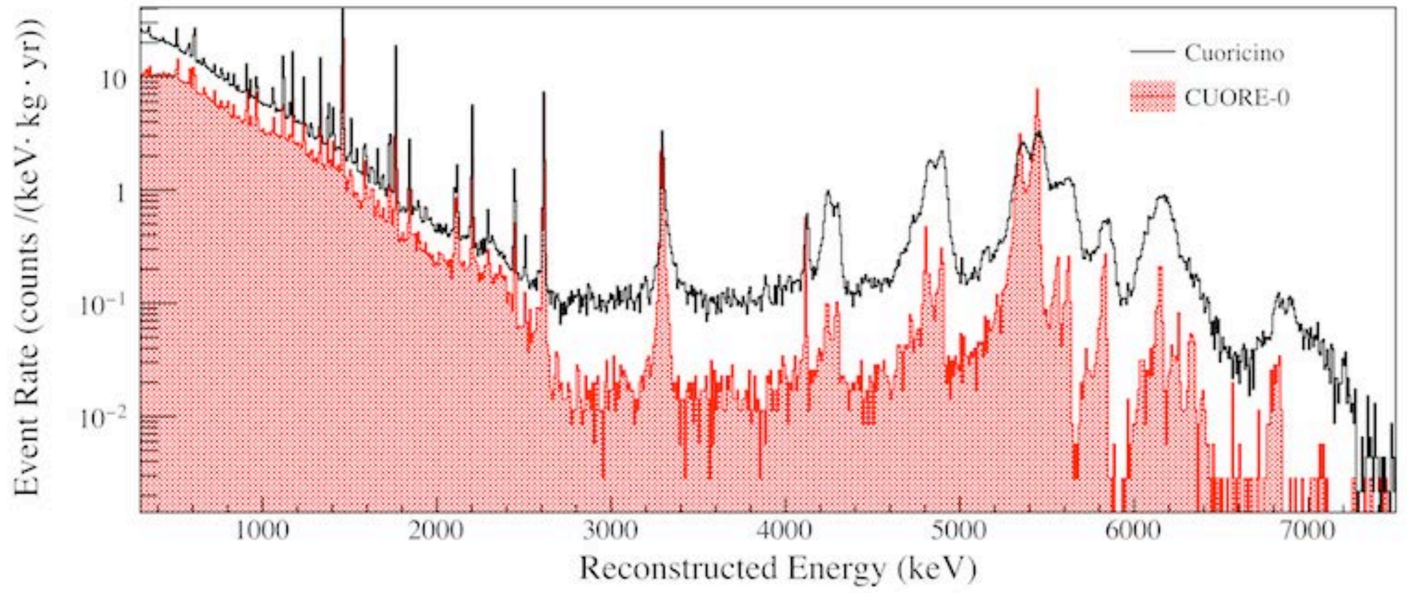
Simulation of detector and cuts





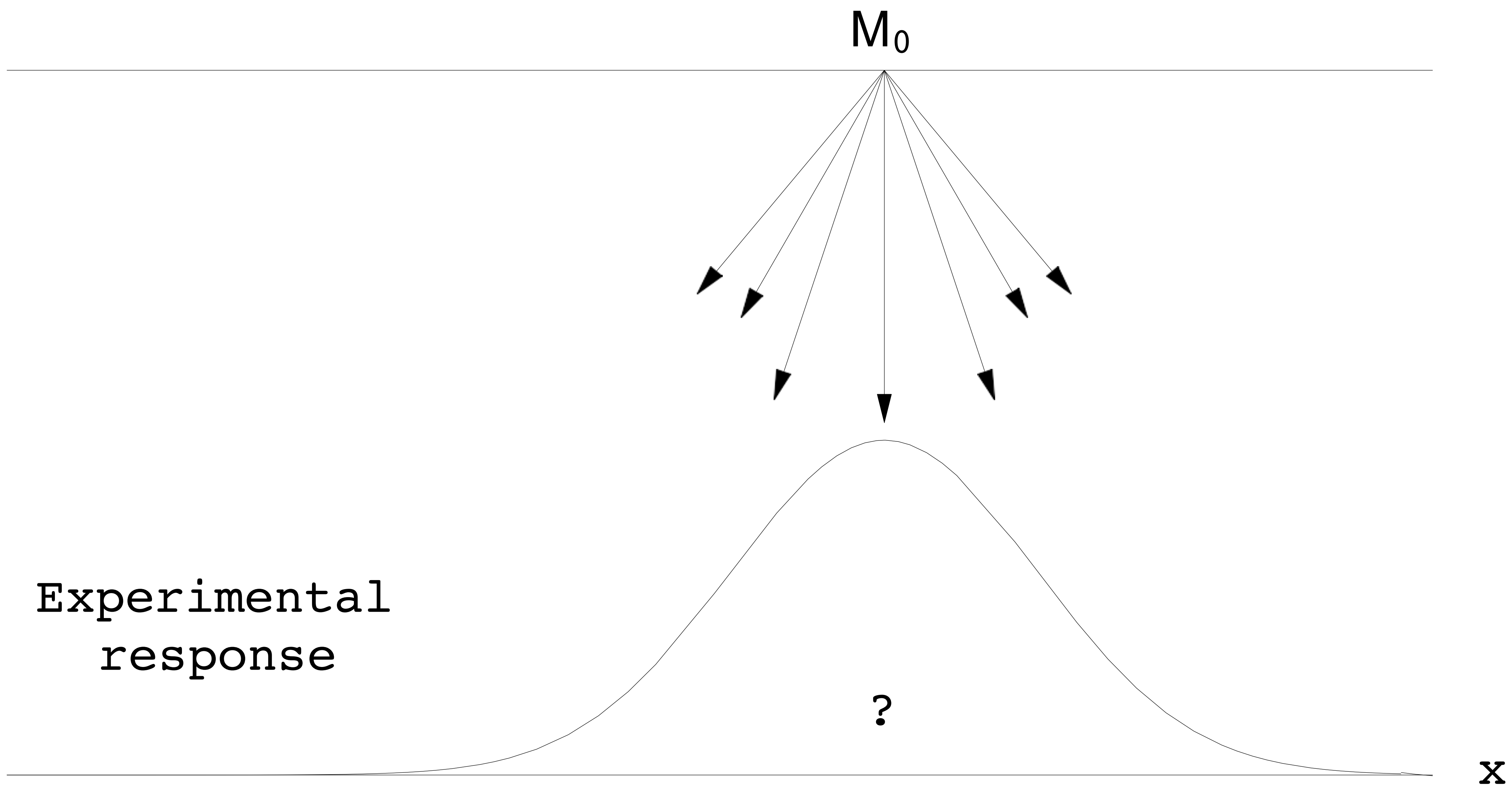
Data selection



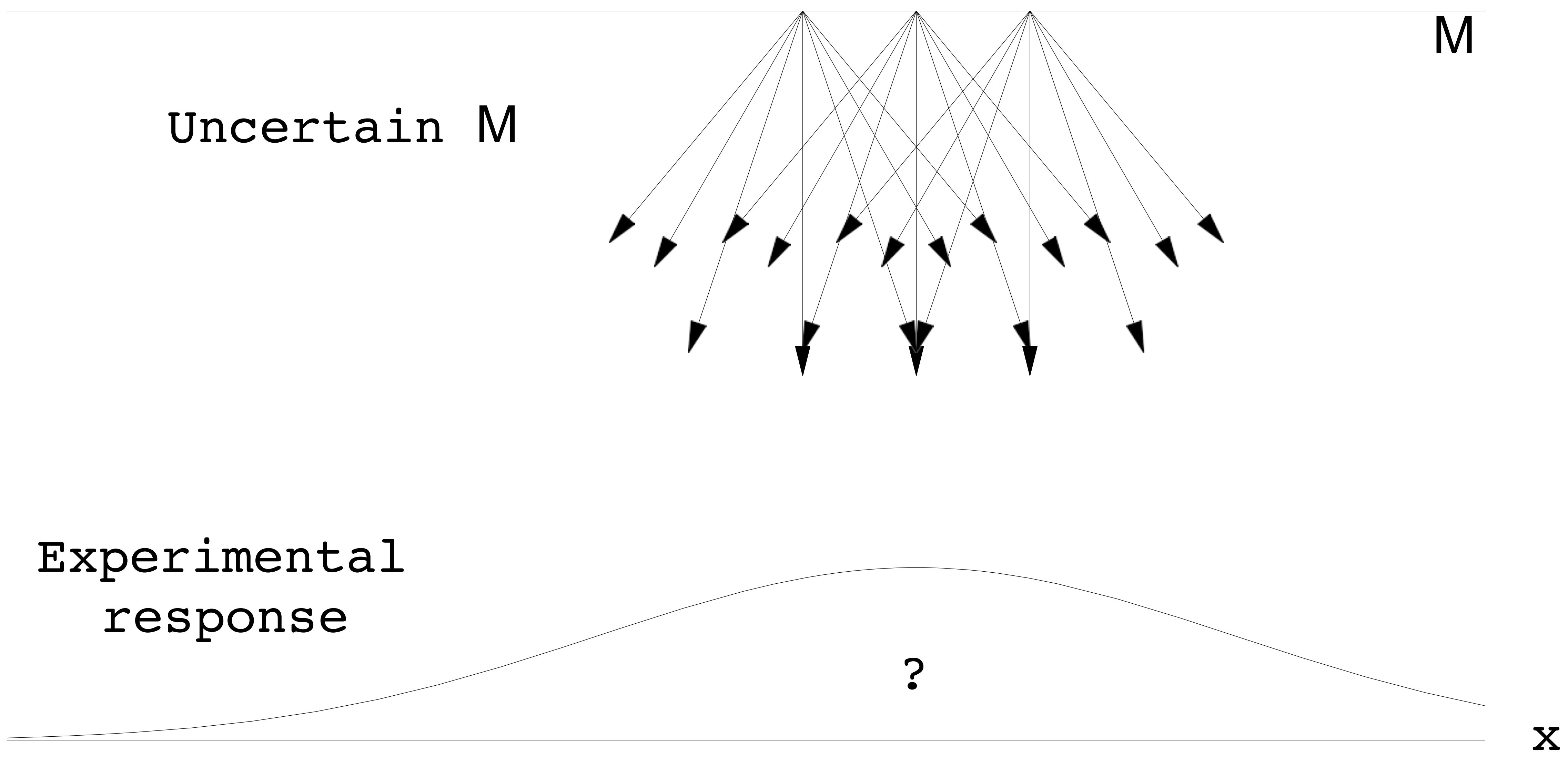


From theory to measurement

- First observation: resulting data are sparse
- Measurements are uncertain!

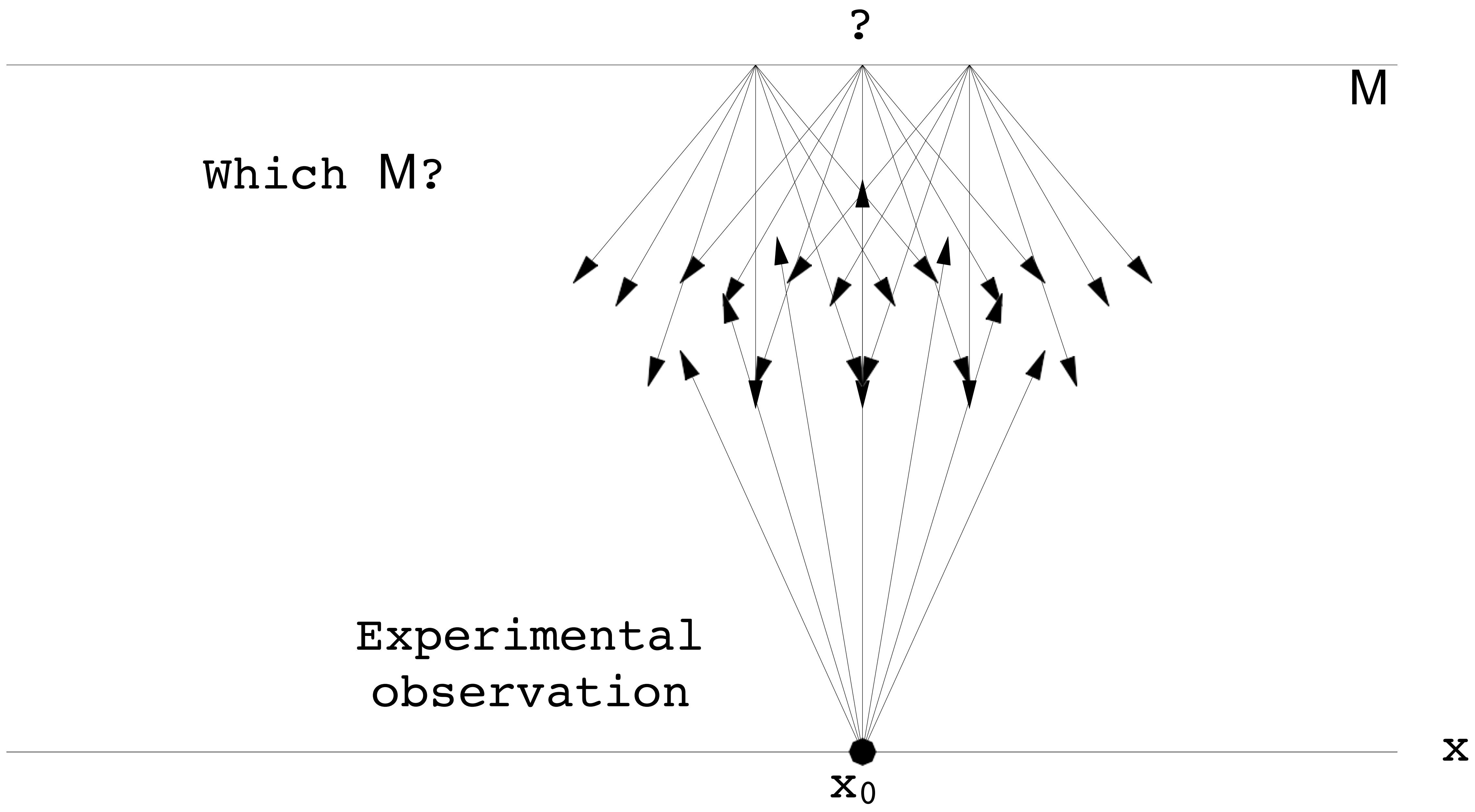


Given μ (exactly known) we are uncertain about x

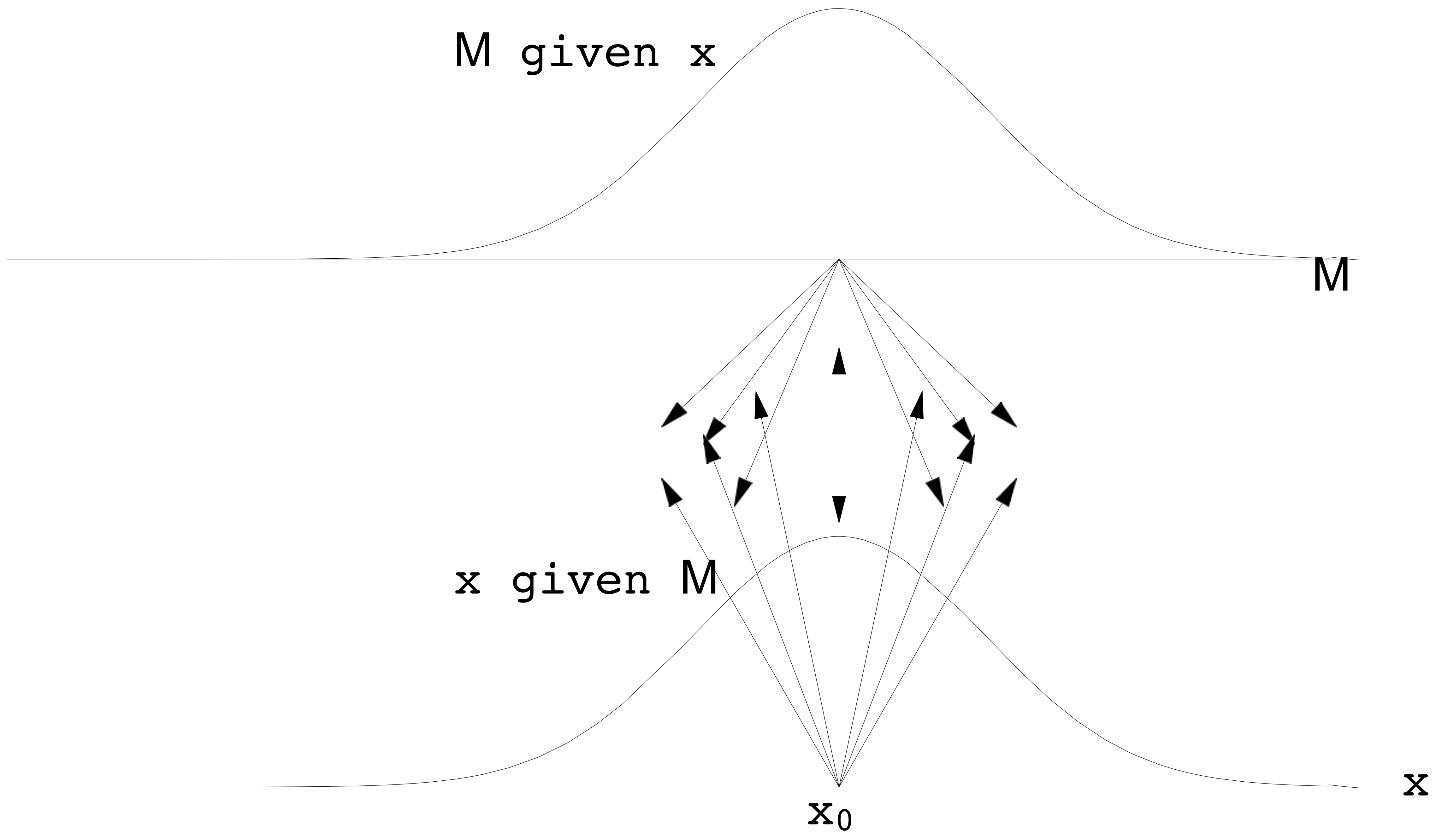


Uncertainty about μ makes us more uncertain about x

Inference



- The observed data is certain: → ‘true value’ uncertain.
- Where does the observed value of x comes from?

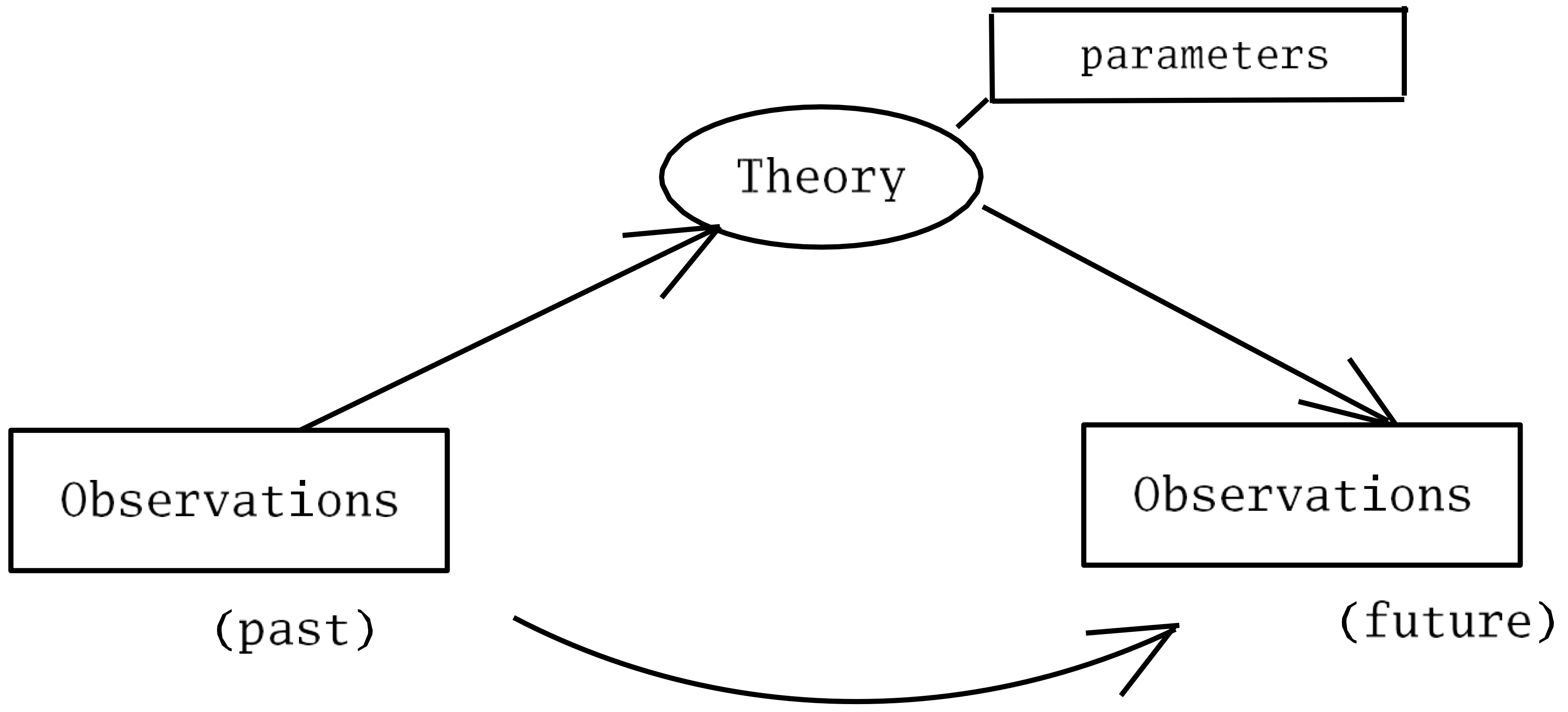


- Given x, μ is uncertain

Measurements goal

Our (physicists) task is to

- Describe/understand the physical world ⇒ inference of laws and their parameters
- Predict observations ⇒ forecasting



Uncertainty:

- All measurements are affected by uncertainty
- Measurements outcome is unpredictable

Uncertainty and probability:

- Physicists consider absolutely natural and meaningful statements like

  $P(m_{\nu_e}) < 1.1$ eV = 90%

  $P(\tau_{1/2}^{\beta\beta0\nu}(^{76}Ge)) < 10^{26}$ y =90%

  …
- however such statements are considered blaspheme to statistics gurus

Indirect probability

Inference:

- Given the past observations, in general we are not sure about the theory parameter (and/or the theory itself)
- Even if we were sure about theory and parameters, there could be internal (e.g. Q.M.) or external effects (initial/boundary conditions, 'errors', etc) that make the forecasting uncertain.

Inference

- Observe events and gather a set of physical quantities: temperatures, voltages, …
  - ➡ Extract particle properties: momenta, nature, …
- Compare observed distributions to predictions of theory
  - ➡ Estimate the free parameters of the theory
- Quantify the uncertainty in the estimates
  - ➡ Assess level of agreement between observed data and a given theory

Issues: small numbers, noise, backgrounds, uncertainties …

We need a clear definition of PROBABILITY and a coherent scheme

Statistics

“A branch of applied mathematics concerned with the collection and interpretation of quantitative data and the use of probability theory to estimate population parameters”

- Inferential aspect is enhanced …
- Though we (physicists) are usually not interested in population parameters, but rather on physics quantities, theories, and so on.

Inference: learning about theoretical objects from experimental observations

Probability theory:

- Essentially OK when we only consider the mathematical apparatus.
- Inference: messy
- Traditionally, a collection of ad hoc prescriptions

  . . . accepted more by authority than by full awareness of what they mean
- We get often confused between good sense and statistics education

Probability

Probability basics

Set theory
(Venn diagrams)

Consider a set *S* with subsets *A*, *B*, ...

$$\boxed{\begin{array}{c} \text{For all } A \subset S, P(A) \geq 0 \\ P(S) = 1 \\ \text{If } A \cap B = \emptyset, P(A \cup B) = P(A) + P(B) \end{array}}$$

Kolmogorov axioms (1933)
valid in all schemes

Define conditional probability of *A* given *B*:

$$\boxed{P(A|B) = \frac{P(A \cap B)}{P(B)}}$$

Define subsets *A*, *B* independent if:

$$P(A \cap B) = P(A)P(B)$$

$$P(A|B) = \frac{P(A)P(B)}{P(B)} = P(A)$$

Probability schemes (interpretations)

I. Relative frequency

A, B, ... are outcomes of a repeatable experiment

$$P(A) = \lim_{n \to \infty} \frac{\text{times outcome is } A}{n}$$

II. Subjective probability

A, B, ... are hypotheses (statements that are true or false)

$$P(A) = \text{degree of belief that } A \text{ is true}$$

- Both interpretations consistent with Kolmogorov axioms.
- Both have pro and cons: frequency interpretation largely diffused, but subjective probability can provide more natural treatment of non-repeatable phenomena (systematic uncertainties, unfolding,…)

Frequentist Statistics

In frequentist statistics, probabilities are associated only with data, i.e., outcomes of repeatable observations

Probability = limiting frequency

Therefore expressions (probabilities) like

- ➡ P (ββ0ν exists),
- ➡ P ($m(\nu_e) < 1.1$ eV)
- ➡ P(…)

etc. are either 0 or 1, but we don’t know which.

Frequentist statistics tell us what to expect, under the assumption of certain probabilities, about hypothetical repeated observations.

A hypothesis is preferred if the data are found in a region of high predicted probability

Bayes' theorem

From the definition of conditional probability we have:

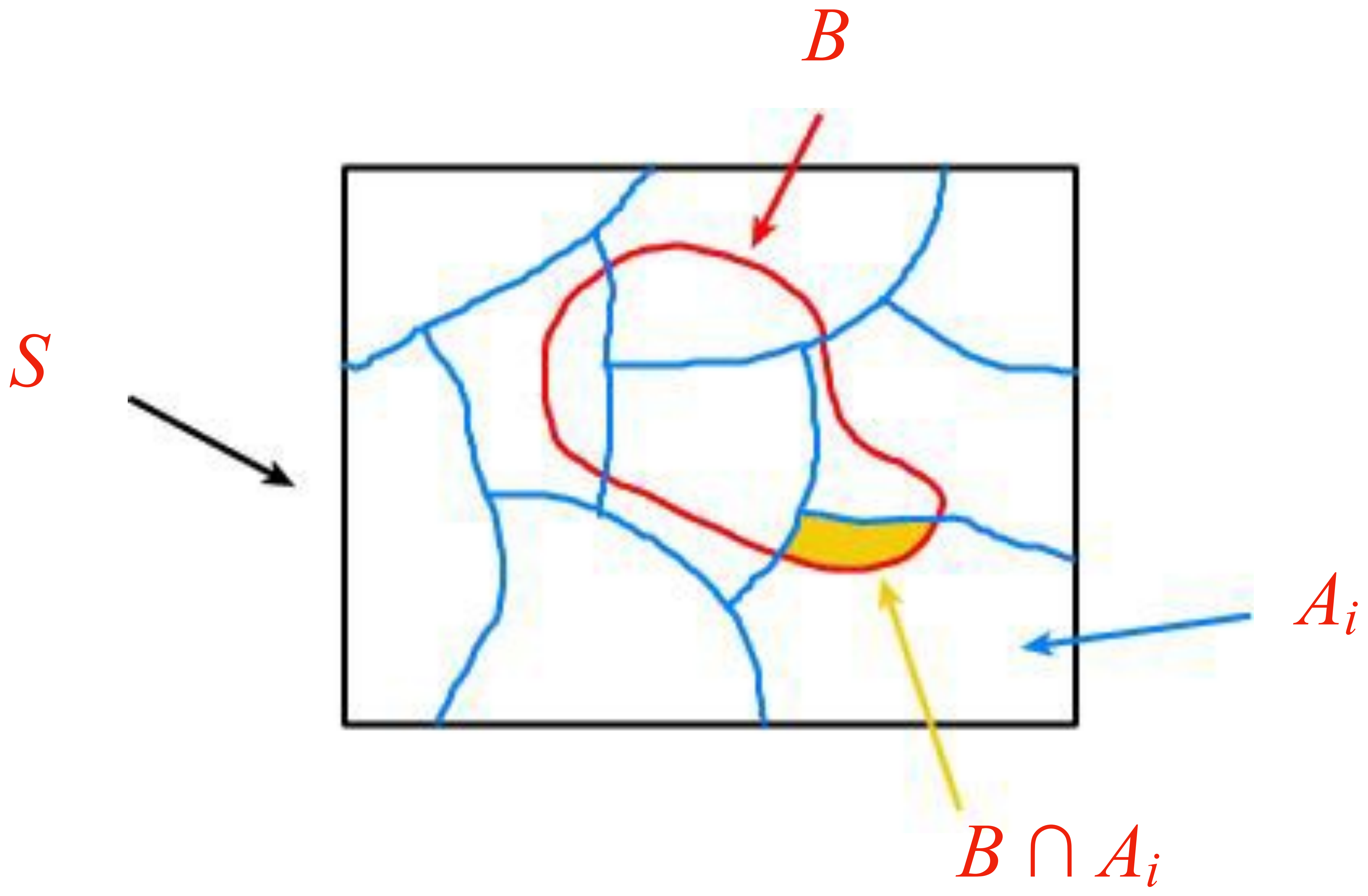
$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad P(B|A) = \frac{P(A \cap B)}{P(A)}$$

but $P(A \cap B) = P(B \cap A)$ so

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$

- First published (posthumously) by the Reverend Thomas Bayes (1702−1761)
- *An essay towards solving a problem in the doctrine of chances,* Philos. Trans. R. Soc. 53 (1763) 370; reprinted in Biometrika, 45 (1958) 293.

Total probability



Consider:

- a subset B of the sample space S
- divided into disjoint subsets A_i such that $\cup_i A_i = S$

Then

$$B = B \cap S = B \cap (\cup_i A_i) = \cup_i (B \cap A_i)$$

$$P(B) = P(\cup_i (B \cap A_i)) = \sum_i P(B \cap A_i)$$

$$P(B) = \sum_i P(B \mid A_i) P(A_i)$$

law of total probability

Bayes' theorem becomes

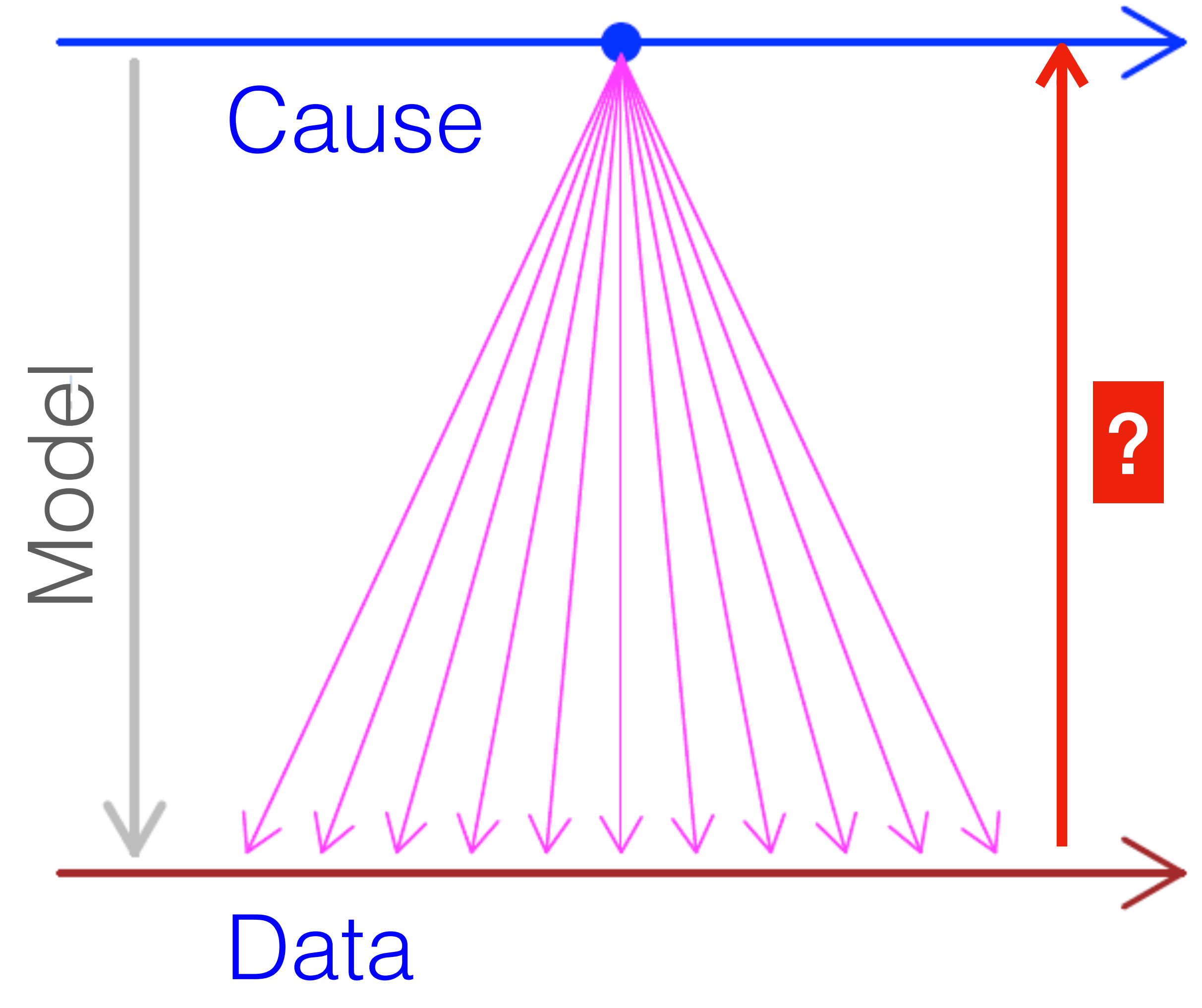
$$\boxed{P(A \mid B) = \frac{P(B \mid A) P(A)}{\sum_i P(B \mid A_i) P(A_i)}}$$

Bayesian approach

Bayes theorem is the ideal tool to manage measurement inversion problem:

- from effect to cause

$$\boxed{P(cause\,|\,data) = \frac{P(data\,|\,cause)P(cause)}{P(data)}}$$



Inversion is straightforward but:

- priors needed (credibility statements)
- priors can be updated

Let’s avoid philosophical discussions here …

Bayesian Statistics – general approach

In Bayesian statistics, use subjective probability for hypotheses:

probability of the data assuming hypothesis *H* (the likelihood)

prior probability, i.e., before seeing the data

$$P(H|\vec{x}) = \frac{P(\vec{x}|H)\pi(H)}{\int P(\vec{x}|H)\pi(H)\,dH}$$

$$P(H|x) \propto P(x|H)P(H)$$

posterior probability, i.e., after seeing the data

normalization involves sum over all possible hypotheses

- Bayes' theorem has an "if-then" character:
  If your prior probabilities is π(H), then you can get how probabilities change in the light of the data.
- No general prescription for priors (subjective!)

A classical example

Just to avoid referring always to COVID-19

- Let’s consider the anti-terrorism tests carried out in the airports and let’s suppose terrorists are carrying explosives (T).
- The test is very effective but not perfect:

$$P(+\,|\,T) = 1 \qquad P(+\,|\,\bar{T}) = 0.02 \qquad P(-\,|\,\bar{T}) = 0.98$$

- Now suppose your test is positive.
- What is the probability that you are (considered) a terrorist?

If we do not have any statement about terrorists, no answer is possible (or better most probably you are arrested!)

However a credible guess is that: $P(T) = 0.001$ and $P(\bar{T}) = 0.999$

Then applying Bayes Theorem we get: $\dfrac{P(T|+)}{P(\bar{T}|+)} = \dfrac{P(+\,|\,T)P(T)}{P(+\,|\,\bar{T})P(\bar{T})} = \dfrac{1 \cdot 0.001}{0.02 \cdot 0.999} \sim 5\,\%$

Conclusion

The lesson is that if the alternative hypothesis has a very small confidence … $P(T)$

… data found in a region of high predicted probability can be irrelevant $P(+|T)$

However for this we need priors (subjective choice)

Distributions

Random variables and probability density functions

- A random variable is a numerical characteristic assigned to an element of the sample space; can be discrete or continuous.

- Suppose outcome of experiment is a continuous value x

$$P(x \text{ found in } [\mathrm{x}, \mathrm{x} + \mathrm{dx}]) = \mathrm{f(x)dx}$$

f(x) is named probability density function (pdf) and $\int_{-\infty}^{+\infty} f(x)dx = 1$

- For discrete outcome x$_i$ (e.g. i = 1, 2, …) we have
  - $P(x_i) = p_i$ - probability mass function
  - $\sum_i P(x_i) = 1$ - x must take on one of its possible values

Types of probability densities

- Outcome of experiment characterized by several values, e.g. an n-component vector, (x_1, ... x_n)

joint pdf: $f(x_1, \dots, x_n)$

- Sometimes we want only pdf of some (or one) of the components

marginal pdf: $f(x_1) = \int \dots \int f(x_1, \dots, x_n) dx_2 \dots dx_n$

x1, x2 independent if $f(x_1, x_2) = f(x_1)f(x_2)$

- Sometimes we want to consider some components as constant

conditional pdf: $g(x_1 \,|\, x_2) = \dfrac{f(x_1, x_2)}{f(x_2)}$

Expectation values

- Consider continuous random variable x with pdf f (x).
- Define expectation (mean) value as

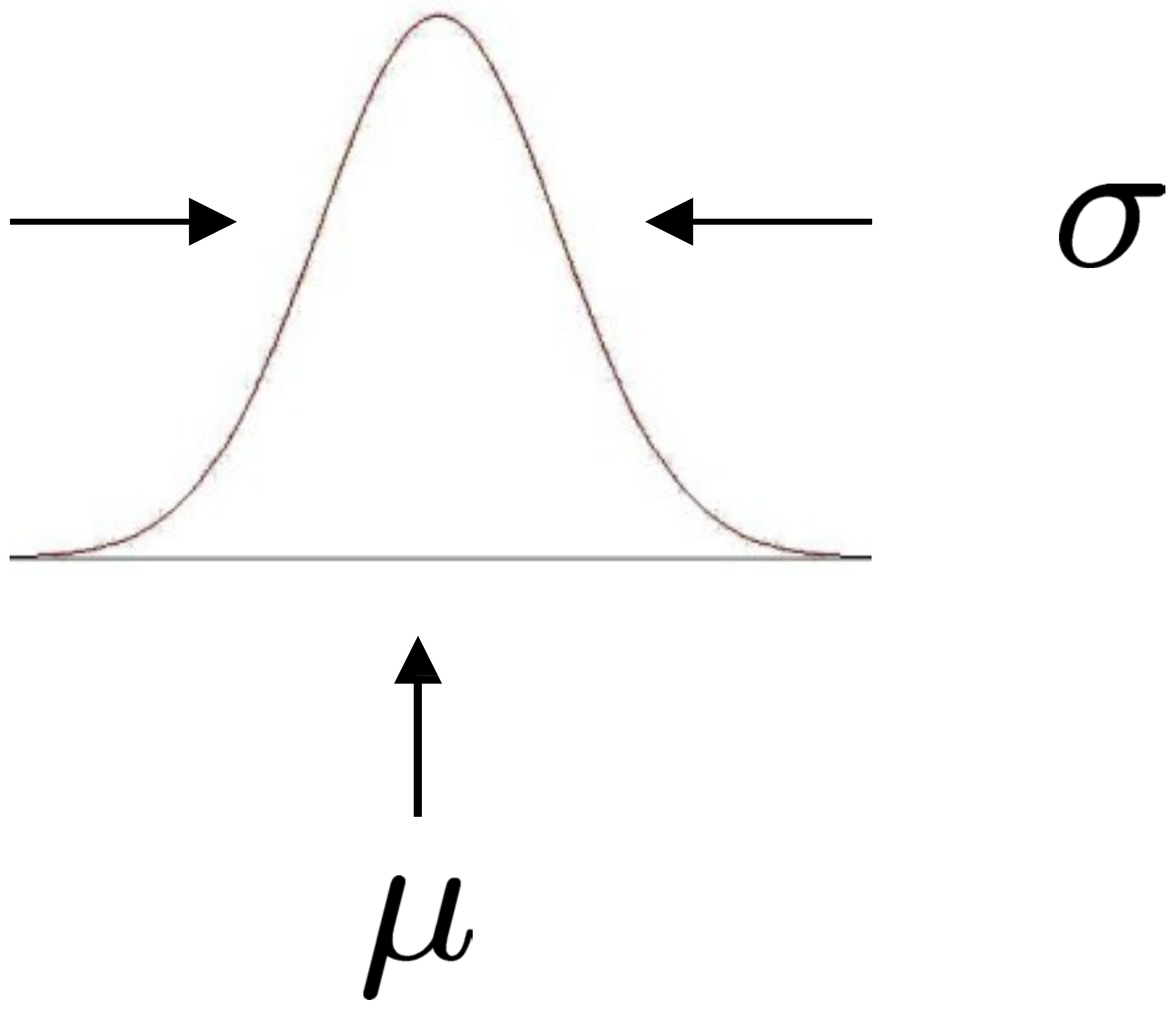
$$E[x] = \int x f(x) dx \quad \rightarrow \quad E[x] = \mu$$ (pdf centre of gravity" of pdf)

- Note: for a function y(x) with pdf g(y):

$$E[y] = \int y g(y) dy = \int y(x) f(x) dx$$

- Variance: $V[x] = E[x^2] - \mu^2 = E[(x-\mu)^2] \quad \rightarrow \quad V[x] = \sigma^2$

- Standard deviation: $\sigma = \sqrt{\sigma^2}$. σ ~ width of pdf, same units as x.



Covariance and correlation

- Define covariance cov[x,y] (also use matrix notation V_{xy}) as

$$\text{cov}[x,y] = E[xy] - \mu_x\mu_y = E[(x-\mu_x)(y-\mu_y)]$$

- Correlation coefficient (dimensionless) defined as

$$\rho_{xy} = \frac{\text{cov}[x,y]}{\sigma_x\sigma_y}$$

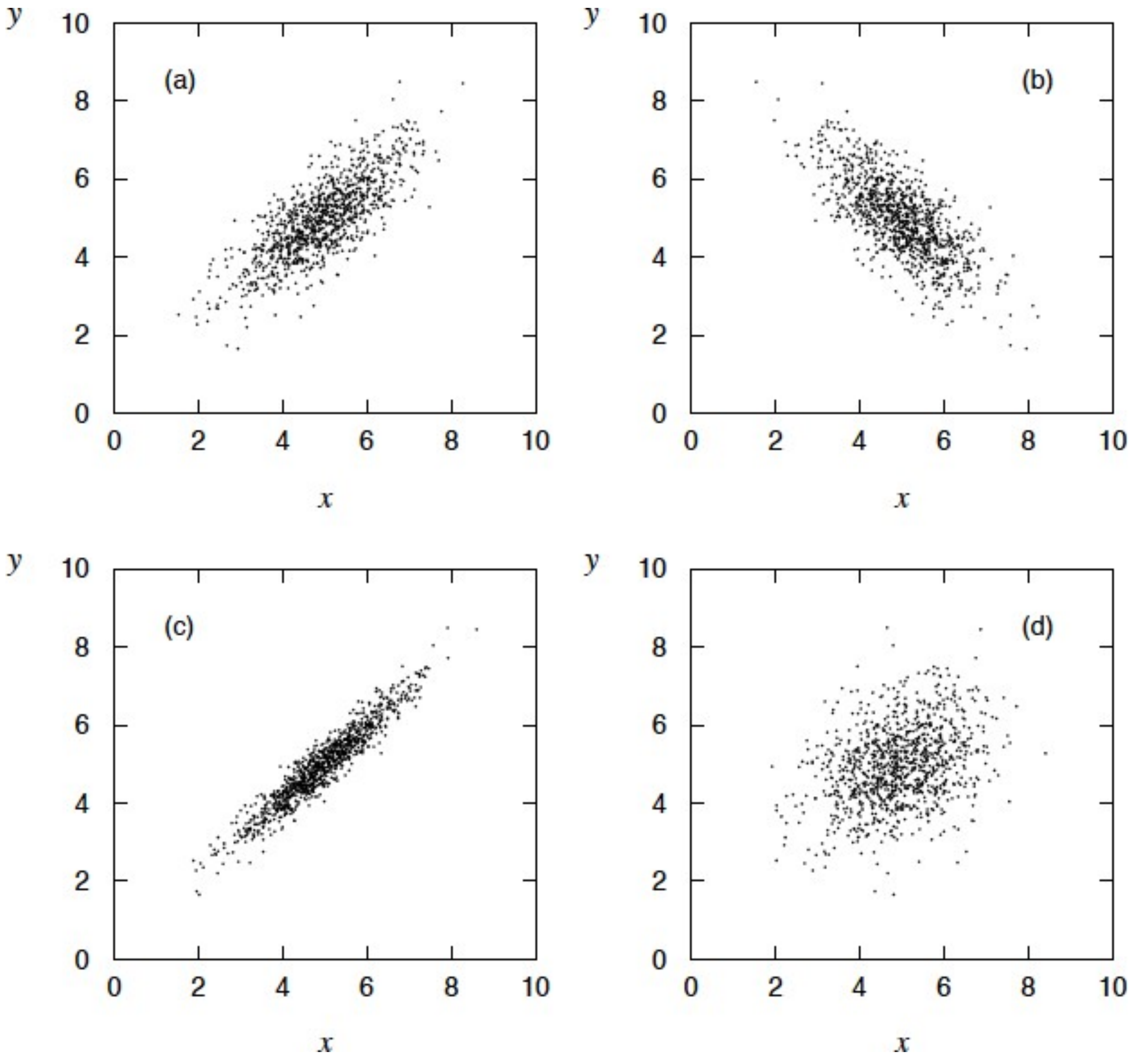
- If x, y, independent (i.e. $f(x,y) = f_x(x)f_y(y)$) then

$$E[xy] = \int\int xy\, f(x,y)\, dx dy = \mu_x\mu_y$$

$\text{cov}[x,y] = 0$ → x and y, ‘uncorrelated’

- Note: converse not always true.

Correlation (cont.)



$\rho = 0.75$

$\rho = -0.75$

$\rho = 0.95$

$\rho = 0.25$

Properties summary

General:

- **Random variable**: integer (usually called r) or real (usually called x)
- **P$_r$ is probability of r** . Dimensionless numbers. $\sum P_r = 1$
- **P(x) is probability density for x**. $[P(x)] = [x]^{-1} or \int P(x)dx = 1$
- **Expectation values**$\langle f \rangle = \sum f(r)P_r or \int f(x)P(x)dx$

Position:

- **Mean**: $\mu = \langle x \rangle$
- **Mode**: $P(\text{mode}) = \max(\text{P(x)})$
- **Median**: $\int^{\text{median}} P(x)dx = 0.5$

Scale:

- $\sigma = \sqrt{\langle (x-\mu)^2 \rangle} = \sqrt{\langle x^2 \rangle - \langle x \rangle^2}$
- **FWHM**=Full Width at Half Max
- Inter-quartile range

Other stuff:

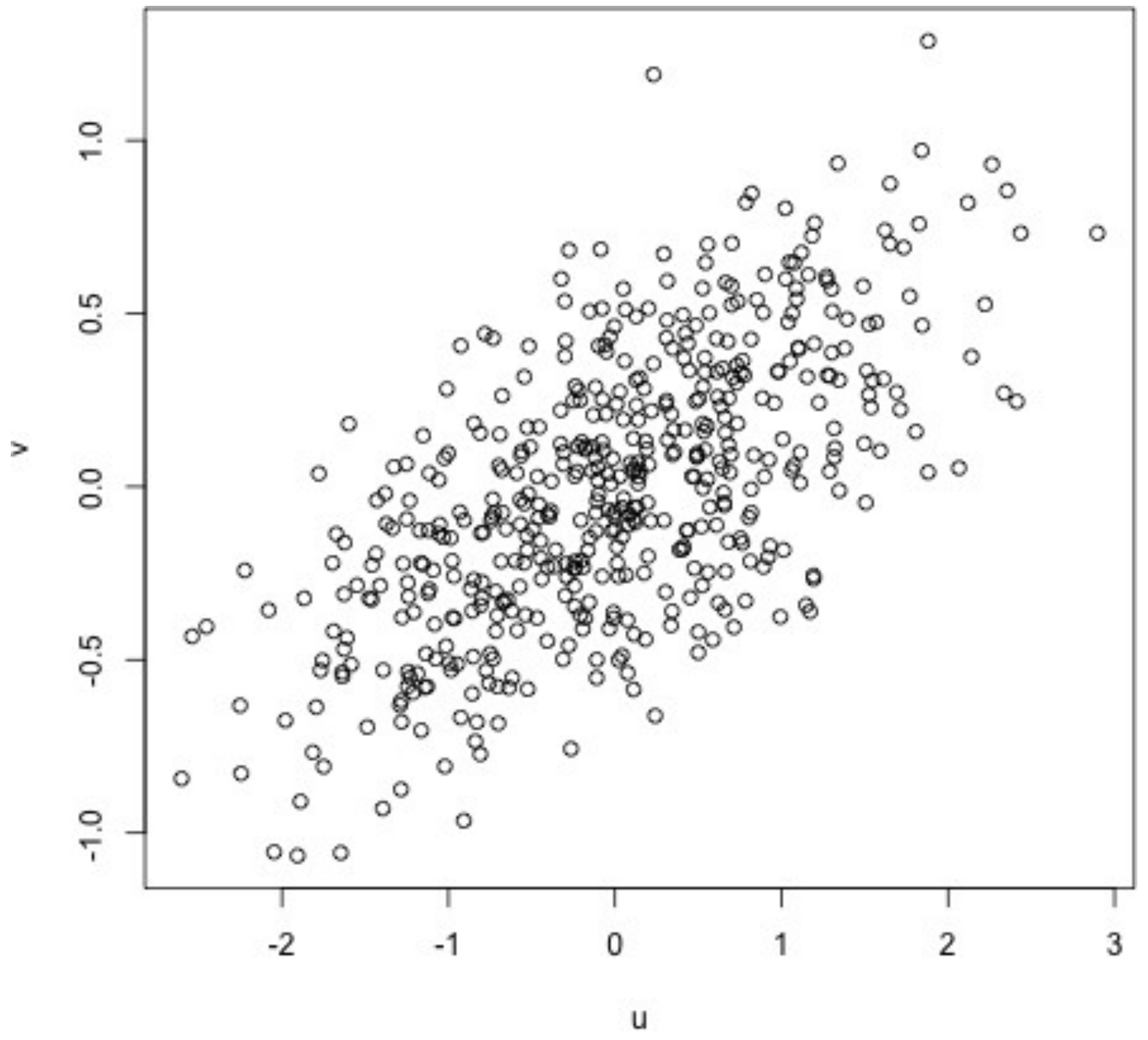
- Skew: $\gamma = \dfrac{\langle (x-\mu)^3 \rangle}{\sigma^3}$
- Kurtosis: $\gamma_2 = \dfrac{\langle (x-\mu)^4 \rangle}{\sigma^4} - 3$
- Moments: $M_N = \langle x^N \rangle \mu_N = \langle (x-\mu)^N \rangle$

Two variables:

- **Covariance**: $Cov(x,y) = \langle xy \rangle - \langle x \rangle \langle y \rangle$
- **Correlation**: $\rho = \dfrac{Cov(x,y)}{\sigma_x \sigma_y}$

Several variables:

- **Covariance**: $C_{ij} = \langle x_i x_j \rangle \langle x_i \rangle \langle x_j \rangle$
- **Correlation**: $\rho_{ij} = \dfrac{C_{ij}}{\sigma_i \sigma_j}$



The Central Limit Theorem

Suppose a random variable x is the sum of several independent identically (or similarly) distributed variables $x_1, x_2, x_3, \ldots, x_N$. Then

(1) The Mean of the distribution for x is the sum of the means: $\mu=\mu_1+\mu_2+\mu_3+\ldots+\mu_N$

(2) The Variance of the distribution for x is the sum of the Variances: $V=V_1+V_2+V_3+\ldots+V_N$

(3) The distribution for x becomes Gaussian for large N

Comments:

- (1) Is obvious … but pay attention to terms: MEAN!
- (2) Is simple and explains 'adding errors in quadrature'
- (1) and (2) do not depend on the form of the distribution
- (3) Explains why Gaussians are 'normal'
- If you find a distribution which is not Gaussian, there must be a reason: probably one contribution dominates
- If a variable has a non-Gaussian pdf you can still apply parts (1) and (2): adding variances, using combination of errors, etc.
- What you can't do is equate deviations with confidence regions (68% within one sigma etc)
- However your variable is probably intermediate and will be a contribution to some final result → Gaussian by (3)

CLT proof

Show that if you convolute P(x) with itself N(→∞) times you get a Gaussian

Given P(x), Fourier Transform is $\tilde{P}(k) = \int P(x)e^{ikx}dx = \langle e^{ikx} \rangle$

Expand and separate: $1 + ik\langle x \rangle + \frac{(ik)^2}{2!}\langle x^2 \rangle + \frac{(ik)^3}{3!}\langle x^3 \rangle \ldots$

Take the logarithm and use $\ln(1+\alpha) = \alpha - \frac{\alpha^2}{2} + \frac{\alpha^3}{3} + \ldots$

Get series in k: $\ln \tilde{P}(k) = (ik)\kappa_1 + \frac{(ik)^2}{2!}\kappa_2 + \frac{(ik)^3}{3!}\kappa_3 + \ldots$

where the κ_r are the expectation values of x^r: $\kappa_r = \langle x^r \rangle$

If x is scaled by factor α, then $\kappa_r \to \alpha^r \kappa_r$

Fact: The FT of a convolution is the product of the individual FTs.

So the log of the FT of a convolution is the sum of the logs and $K_r = N\kappa_r$.

To discuss shape, scale by standard deviation $\sqrt{K_2}$

$K_2' = 1,\ K_r' = K_r/\sqrt{K_2}^{\,r} = N\kappa_r/(N\kappa_2)^{r/2}$, vanishes as N → ∞ for r>2.

So in the large N limit all K_r with r ≥ 3 vanish, and the log of the FT is quadratic

The FT itself is the exponential of a quadratic, i.e. a Gaussian.

Transforming, the (back) FT of a Gaussian is also a Gaussian.

Some distributions

Distribution/pdf	Example use in HEP
Binomial	Branching ratio
Multinomial	Histogram with fixed *N*
Poisson	Number of events found
Uniform	Monte Carlo method
Exponential	Decay time
Gaussian	Measurement error
Chi-square	Goodness-of-fit
Cauchy	Mass of resonance
Landau	Ionization energy loss
Beta	Prior pdf for efficiency
Gamma	Sum of exponential variables
Student’s *t*	Resolution function with adjustable tails

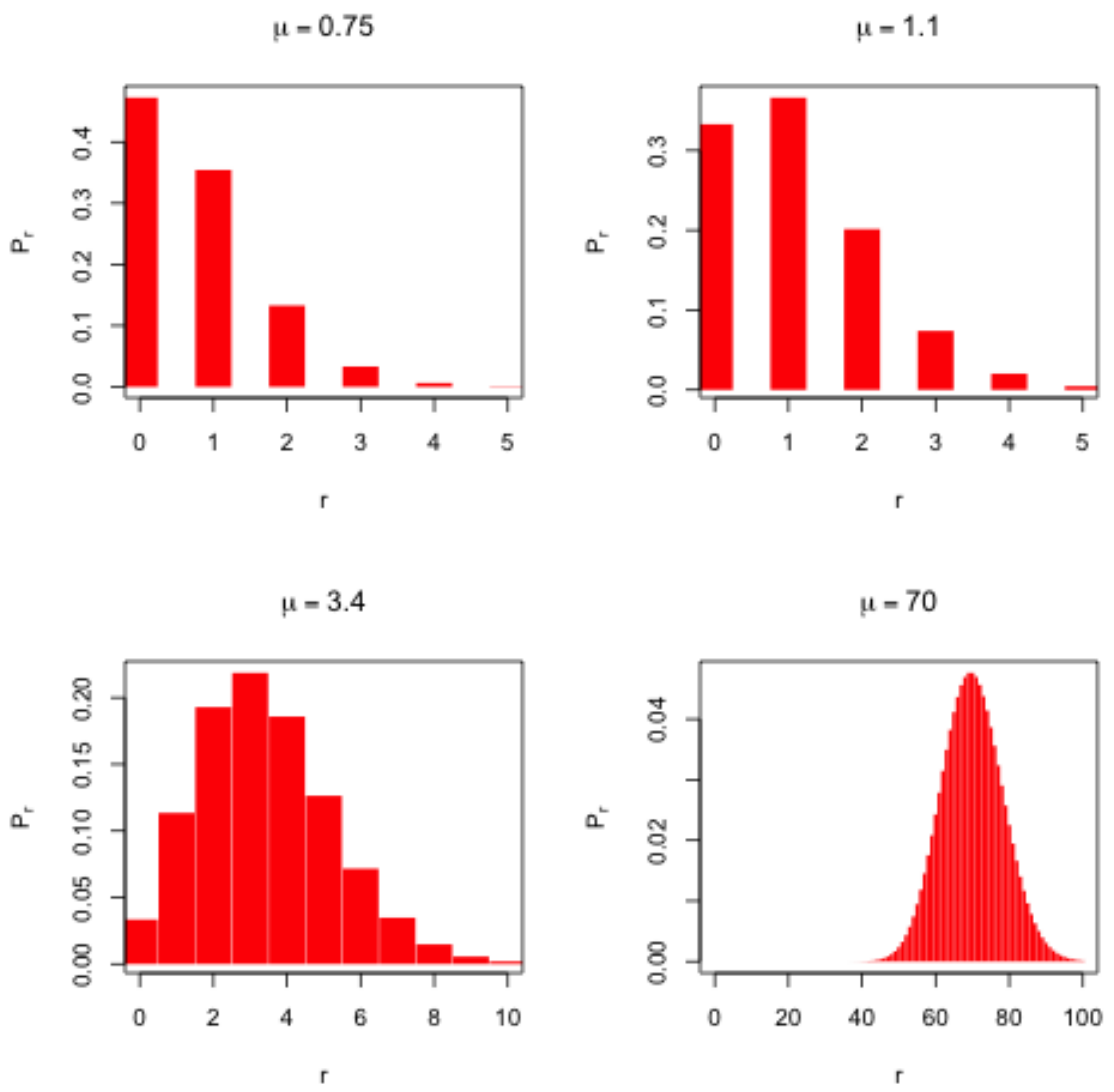
Poisson

Memoryless random source. Mean number μ. Actual number r

$$P(r, \mu) = \frac{e^{-\mu}\mu^r}{r!}$$

Classic examples: Geiger counter clicks. Photomultipliers. Rare decays.
Counterexamples: Photons from lasers. Traffic

- Vital fact: $\sigma = \sqrt{\mu}$
- Small μ : mode is 0
- μ > 1: peak develops
- Distribution has positive skew - tail to high values
- Large μ: shape becomes Gaussian



Important convolutions:

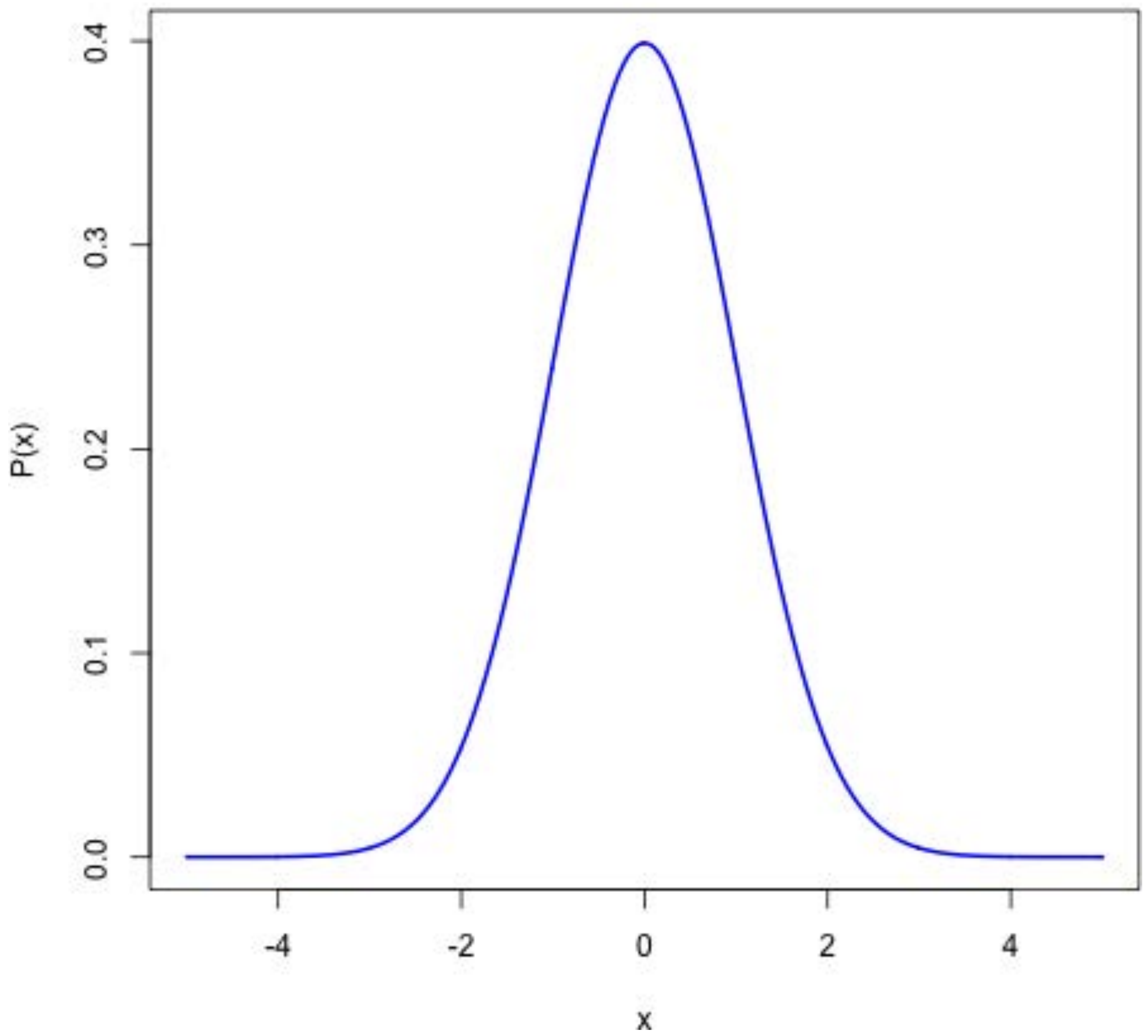
- Poisson * Poisson = Poisson
  Separate sources add and can be treated as a single source
- Poisson * Binomial = Poisson
  A poisson source modified by a binomial detection efficiency gives a poisson number of detected events

Gaussian

Also known as 'normal distribution'

$$P(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

- (Inaccurately) called the 'Bell curve'
- μ is mean and mode and median
- σ is standard deviation
  - 68.27% of area within 1 σ
    so 1/3 of error bars should miss!
  - 95.45% of area within 2 σ
  - 99.73% of area within 3 σ



- Describes: large μ Poisson, measurement errors, height, IQ, ...
- Does not describe: Weight, wealth, ...
- Vital fact: Thanks to Central Limit Theorem: convolution of N random variables P(x) tends to Gaussian for large N, irrespective of P(x).

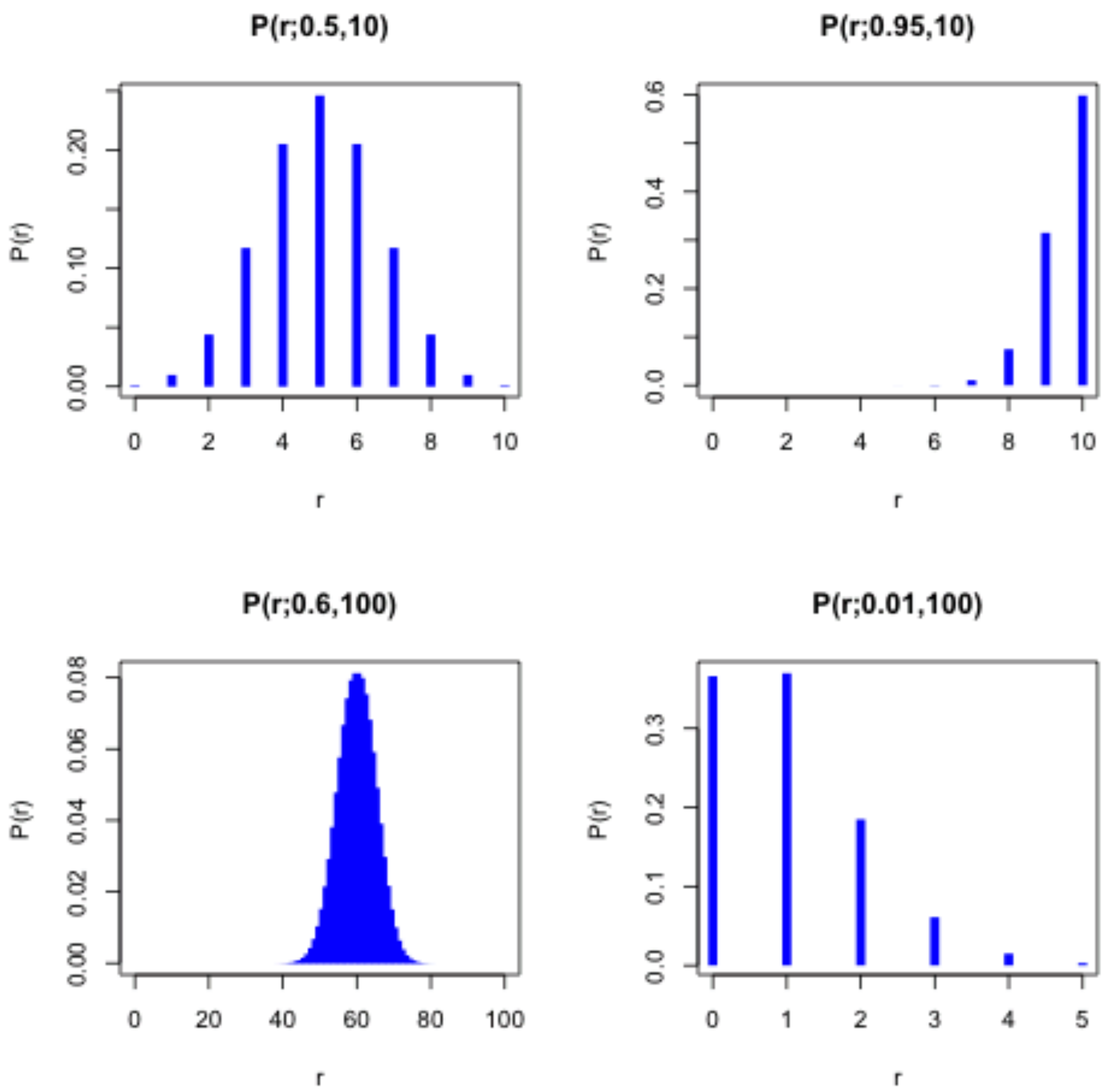
Binomial

Probability of r `successes' from n trials, each with probability p.

$$P(r;n,p) = \binom{n}{r} p^r q^{n-r} = \frac{n!}{r!(n-r)!} p^r q^{1-r}$$

with q=(1-p)

$$\mu = np; \quad \sigma = \sqrt{npq}$$



Limit: n large, p small, np = μ fixed P(r) → Poisson

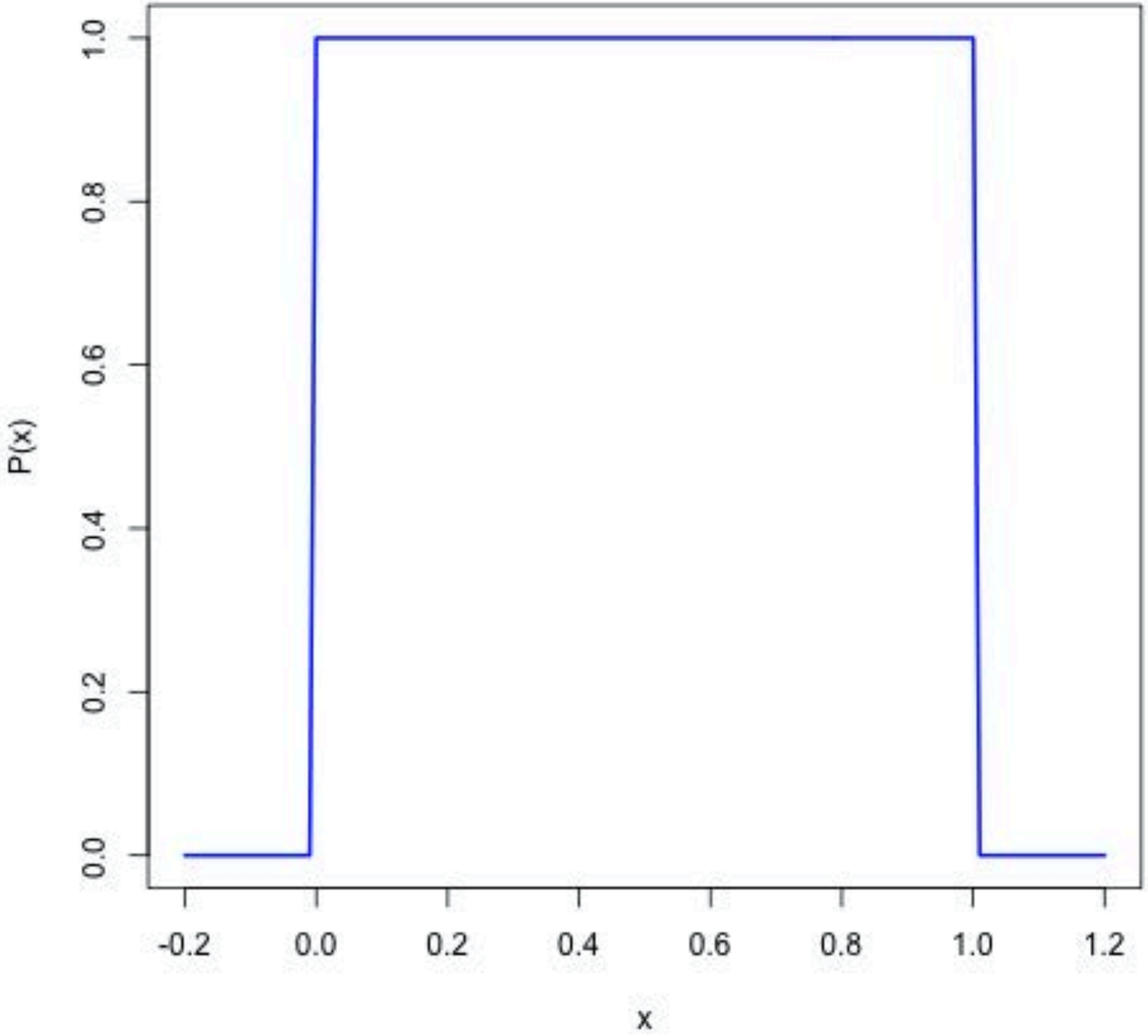
Vital Fact: Basically just like tossing coins

Uniform

Generally $P(x) = \frac{1}{a}$

between $\mu - a/2$ and $\mu + a/2$

Vital fact: Standard Deviation $\sigma = \frac{a}{\sqrt{12}}$



Multinomial distribution

Like binomial but now m outcomes instead of two, probabilities are $\vec{p} = (p_1, \ldots, p_m)\,,$ with $\sum_{i=1}^{m} p_i = 1\,.$

For N trials we want the probability to obtain:

n_1 of outcome 1,
n_2 of outcome 2,
⋮
n_m of outcome m.

This is the multinomial distribution for $\vec{n} = (n_1, \ldots., n_m)$

$$f(\vec{n}; N, \vec{p}) = \frac{N!}{n_1! n_2! \cdots n_m!} p_1^{n_1} p_2^{n_2} \cdots p_m^{n_m}$$

Now consider outcome i as 'success', all others as 'failure' → all n_i individually binomial with parameters N, p_i, for all i

$$E[n_i] = Np_i, \quad V[n_i] = Np_i(1 - p_i)$$

$$V_{ij} = Np_i(\delta_{ij} - p_j)$$

One can also find the covariance to be

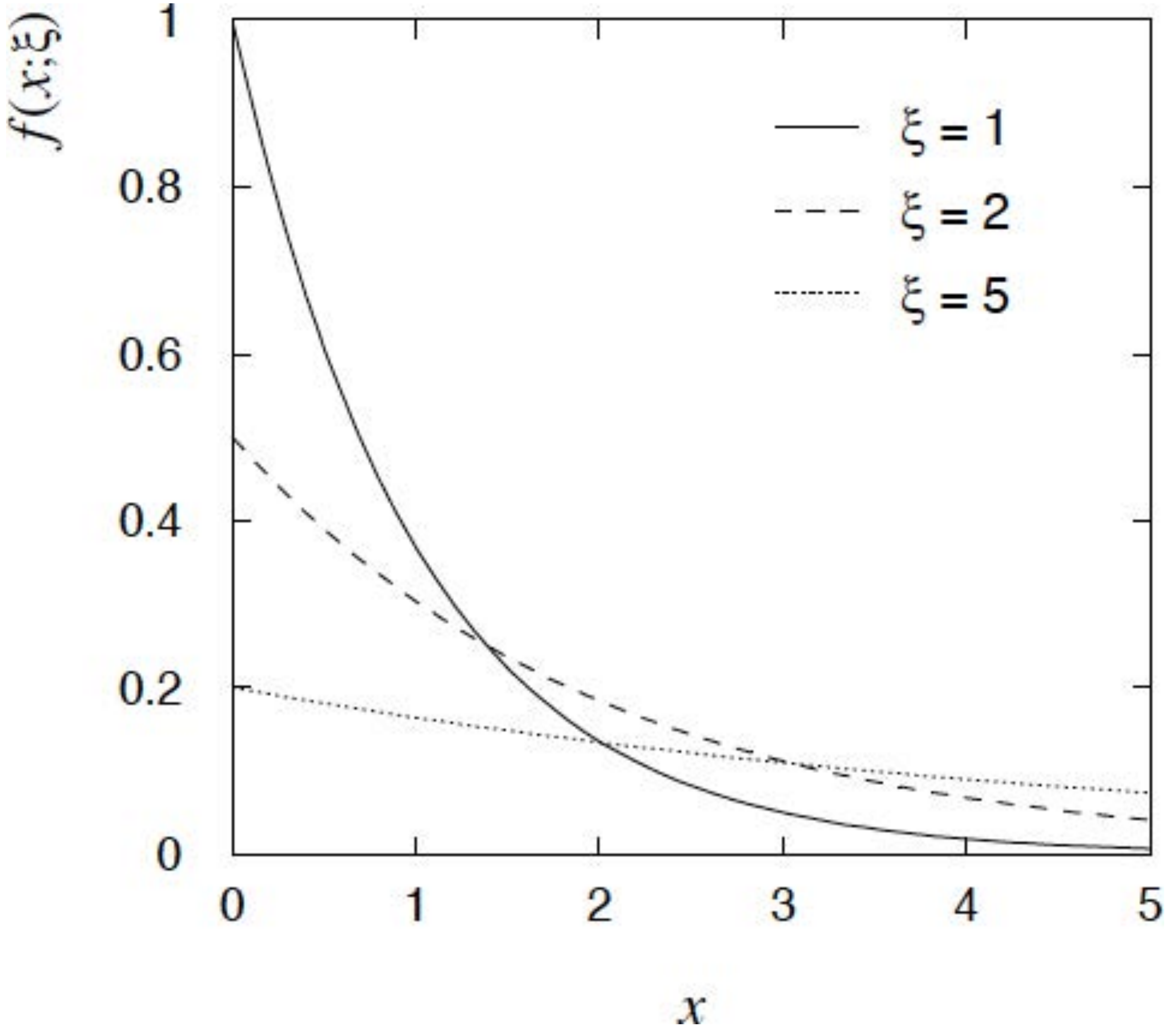
Exponential distribution

The exponential pdf for the continuous r.v. x is defined by:

$$f(x;\xi) = \begin{cases} \frac{1}{\xi} e^{-x/\xi} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$E[x] = \xi$$

$$V[x] = \xi^2$$



Example: proper decay time t of an unstable particle

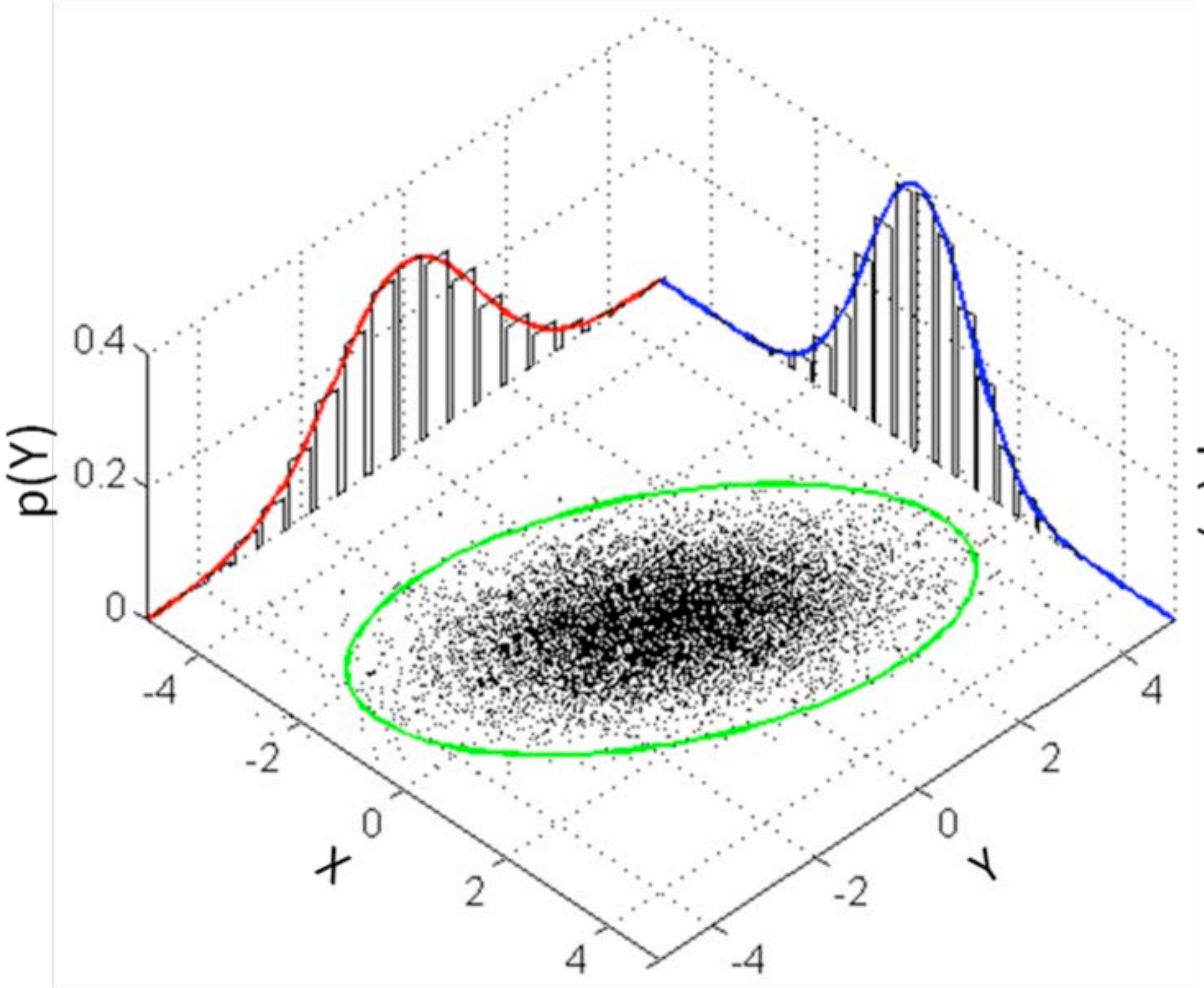
$$f(t;\tau) = \frac{1}{\tau} e^{-t/\tau} \qquad (\tau = \text{mean lifetime})$$

Lack of memory (unique to exponential): $f(t - t_0 | t \geq t_0) = f(t)$

Multivariate Gaussian distribution

Multivariate Gaussian pdf for the vector $\vec{x} = (x_1,\ldots,x_n)$:

$$f(\vec{x};\vec{\mu},V) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})^T V^{-1}(\vec{x}-\vec{\mu})\right]$$



$\vec{x}$, $\vec{\mu}$ are column vectors, and $\vec{x}^T$, $\vec{\mu}^T$ are transpose (row) vectors,

$$E[x_i] = \mu_i, \ , \qquad \mathrm{cov}[x_i, x_j] = V_{ij} \ .$$

For $n = 2$ this is $f(x_1, x_2, ; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$

$$\times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]\right\}$$

where $\rho = \mathrm{cov}[x_1, x_2]/(\sigma_1\sigma_2)$ is the correlation coefficient.

Chi-square (χ^2) distribution

The chi-square pdf for the continuous r.v. z (z ≥ 0) is defined by

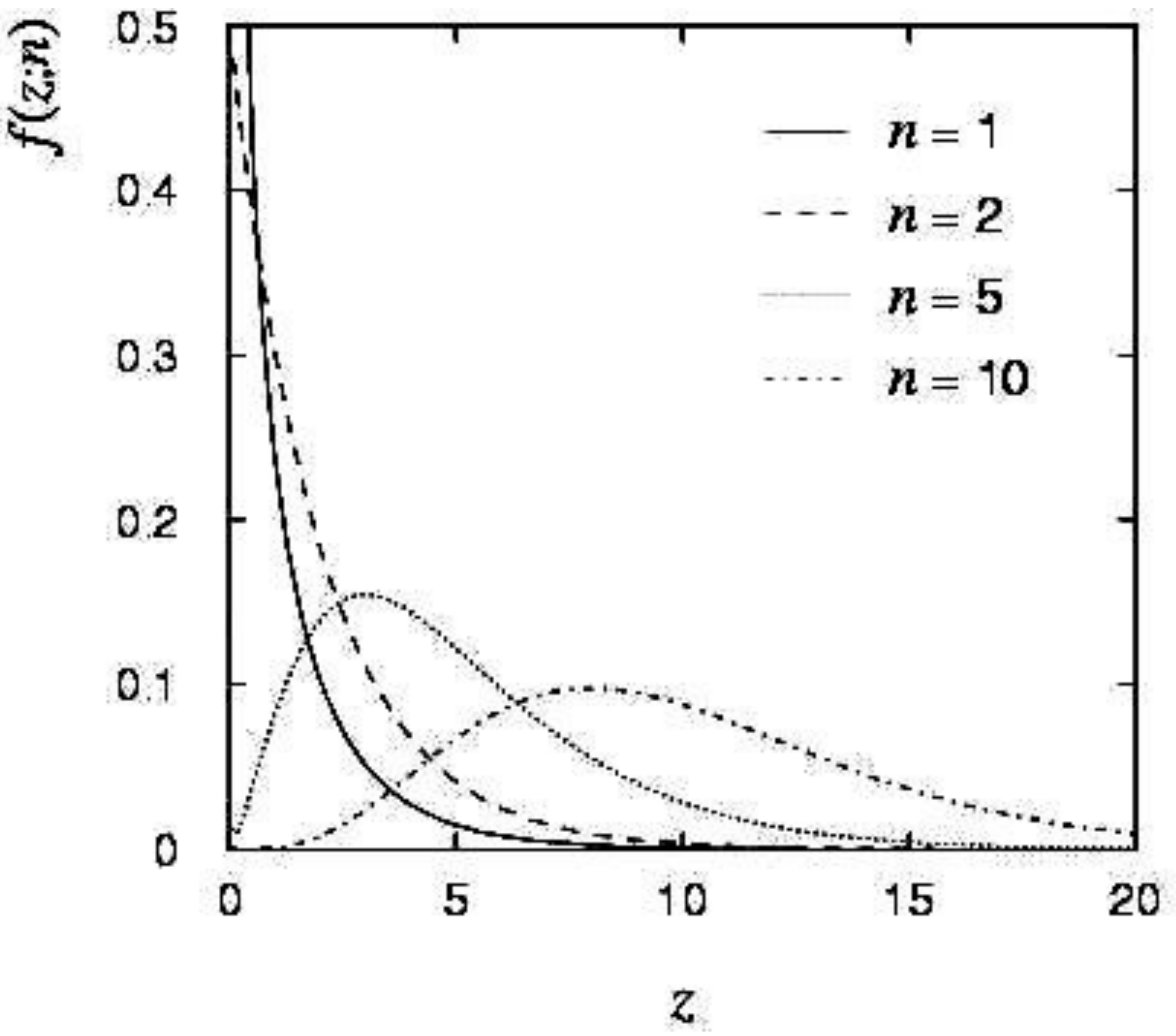
$$f(z;n) = \frac{1}{2^{n/2}\Gamma(n/2)} z^{n/2-1} e^{-z/2}$$

n = 1, 2, ... = number of 'degrees of freedom' (dof)

$$E[z] = n\,, \quad V[z] = 2n\,.$$

For independent Gaussian x_i, i = 1, ..., n, means μ_i, variances σ_i^2,

follows χ^2 pdf with n dof.



Example: goodness-of-fit test variable especially in conjunction with method of least squares.

$$z = \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$

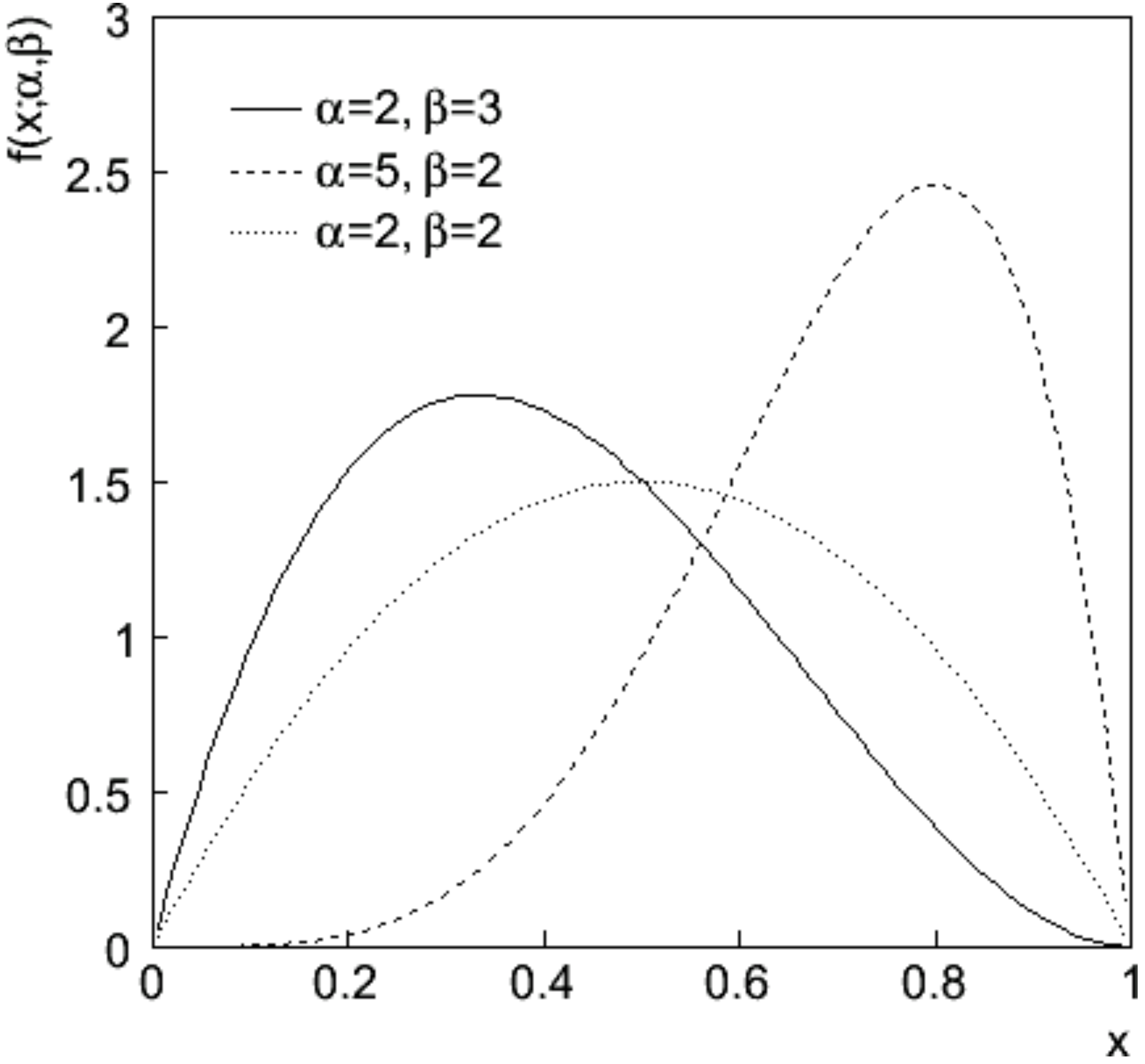
Beta distribution

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}$$

$$E[x] = \frac{\alpha}{\alpha+\beta}$$

$$V[x] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

- Often used to represent pdf of continuous r.v. non-zero only between finite limits
- In Bayesian inference is the conjugate prior probability distribution for the Bernoulli, binomial, negative binomial and geometric distributions



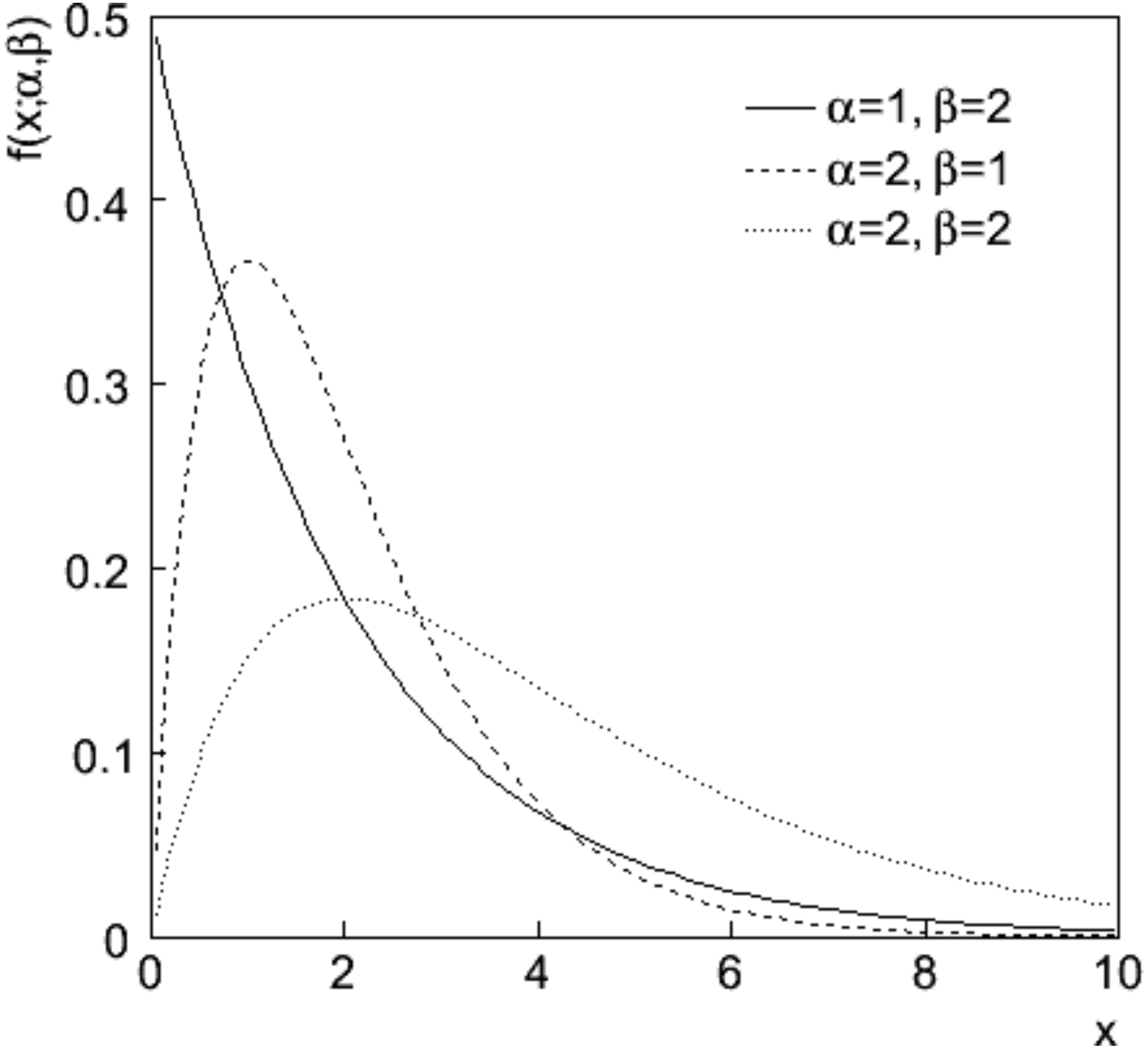
Gamma distribution

$$f(x;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}}\, x^{\alpha-1}\, e^{-x/\beta}$$

$$E[x] = \alpha\beta$$

$$V[x] = \alpha\beta^2$$

- Often used to represent pdf of continuous r.v. nonzero only in [0,∞].
- Also e.g. sum of n exponential r.v.s or time until nth event in Poisson process ~ Gamma



Student's *t* distribution

$$f(x;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}\left(1+\frac{x^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)}$$

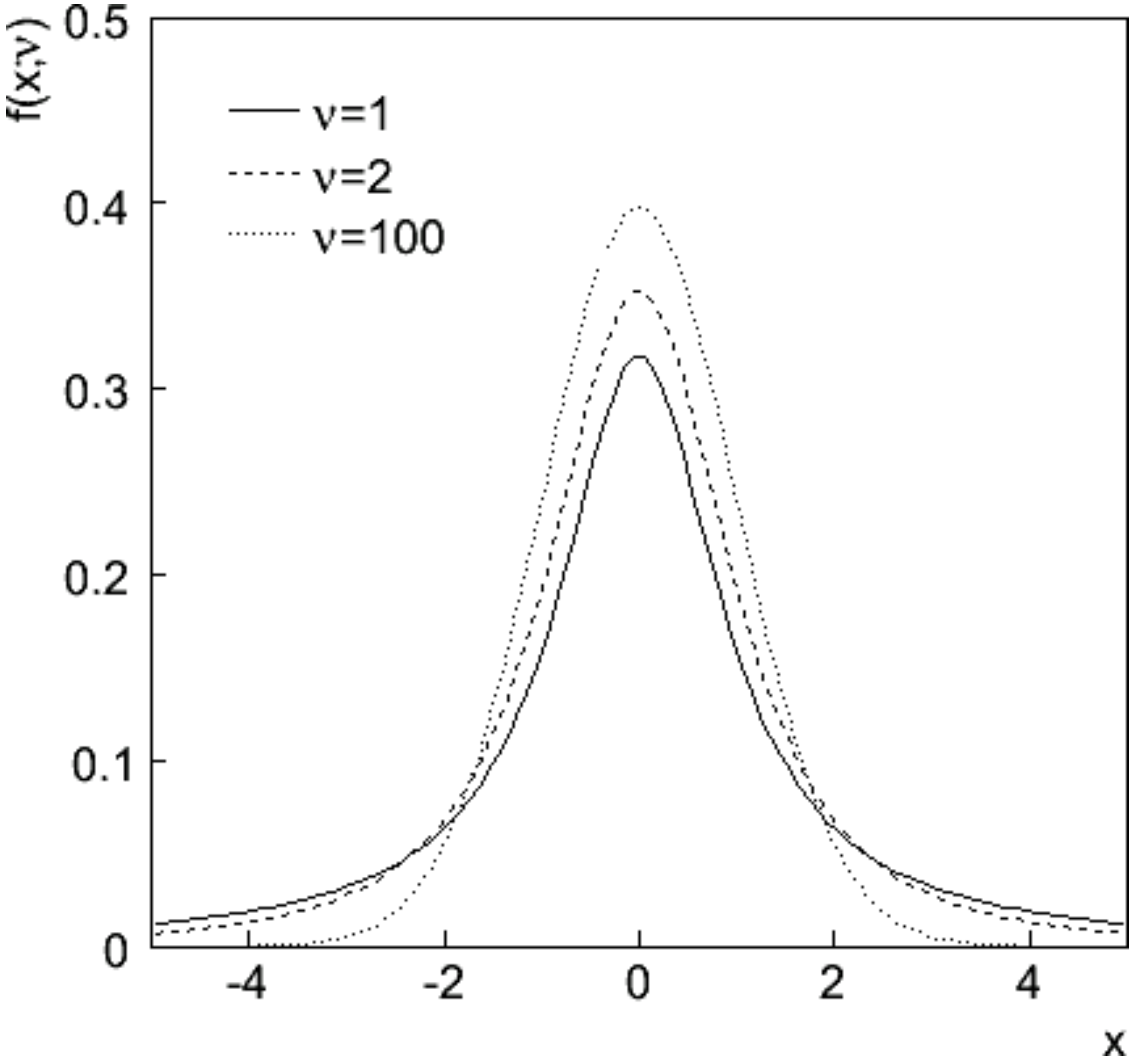
$$E[x] = 0 \quad (\nu > 1)$$

$$V[x] = \frac{\nu}{\nu-2} \quad (\nu > 2)$$

ν = number of degrees of freedom (not necessarily integer)

ν = 1 gives Cauchy (Lorentz),

ν → ∞ gives Gaussian.



Estimators

General properties (1/2)

Assume that:

- We have a sample (x_1,x_2,,,x_n) from a given population
- All parameters of the population are known except some parameter θ.
- We want to determine the unknown parameter θ, from the given observations. In other words we want to determine a number or range of numbers from the observations that can be taken as a value of θ.
- **Estimator:** a method of estimation.
- **Estimate:** a result of an estimator
- **Point estimation:** the estimation of the population parameter with one number

Problem of statistics is not to find estimates but to find estimators

- Estimator is not rejected because it gives one bad result for one sample. It is rejected when it gives bad results in a long run
- Estimator is accepted or rejected depending on its sampling properties
- Estimator is judged by the properties of the distribution of estimates it gives rise.

General properties (2/2)

- Since estimator t gives rise to an estimate that depends on sample points ($x_1,x_2,,,x_n$), the estimate is a function of sample points.

- Sample points are random variables, therefore estimate is random variable and has a probability distribution.

- We want the estimator to have several desirable properties like:
  - Consistency limiting property: convergence to the a fixed value θ_0: $\forall$ small ε,η $\exists$ n_0 s.t. $P(|t_n-\theta_0|<\varepsilon)=1-\eta$ if $n>n_0$
  - Unbiasedness the bias $B_\theta = E(t_n-\theta)=E(t_n)-\theta$ (expectation value of the estimate error) → 0
  - Minimum variance $V(t_n) = E([t_n-E(t_n)]^2)$ minimum variance unbiased estimator are not always possible.

- In general it is not possible for an estimator to have all these properties.

Parameter estimation (frequentist)

Suppose we have a pdf characterized by one or more parameters (e.g.):

$$f(x, \theta) = \frac{1}{\theta} e^{-x/\theta}$$

x = random variable
θ = parameter

Suppose we have a sample of observed values $\overrightarrow{x} = (x_1, \ldots, x_n)$

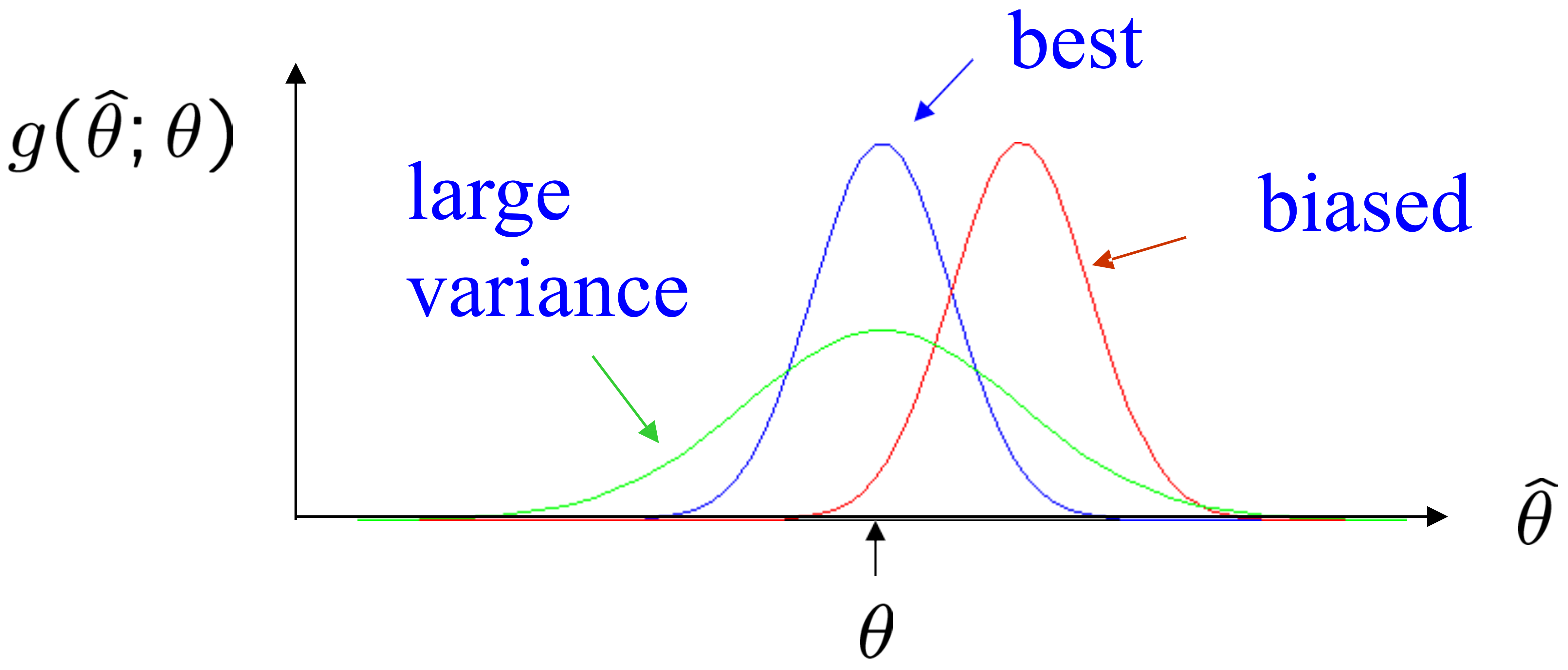
We want to find some function of the data to estimate the (most likely set) of parameter(s):

$$\boxed{\hat{\theta}(\overrightarrow{x})}$$

- We indicate the estimator with a hat
    - ‘estimator’: function of x_1, ..., x_n
    - ‘estimate’: the value of the estimator with a particular data set.

Properties of estimators

Estimators are functions of the data and thus characterized by a sampling distribution with a given (co)variance:



In general they may have a nonzero bias: $b = E[\hat{\theta}] - \theta$

- Desiderata: small variance and small bias
- In general: impossible to optimize with respect to both; trade-off necessary.

Likelihood function and estimator

- Common situation in physics: infinite set of hypotheses described by a continuous (set of) parameter θ

$$f(x\,|\,\theta) \text{ such that } \int f(x\,|\,\theta)d\mathbf{x} = 1$$

- Define the Likelihood function

$$\mathscr{L}(\theta) = \prod_k f(x_k\,|\,\theta)$$

- The most probable value of $\mathscr{L}(\theta)$ is called the maximum-likelihood estimator $\hat{\theta}$.

- The rms (root-mean-square) spread of θ about $\hat{\theta}$ is a conventional measure of the accuracy of the determination

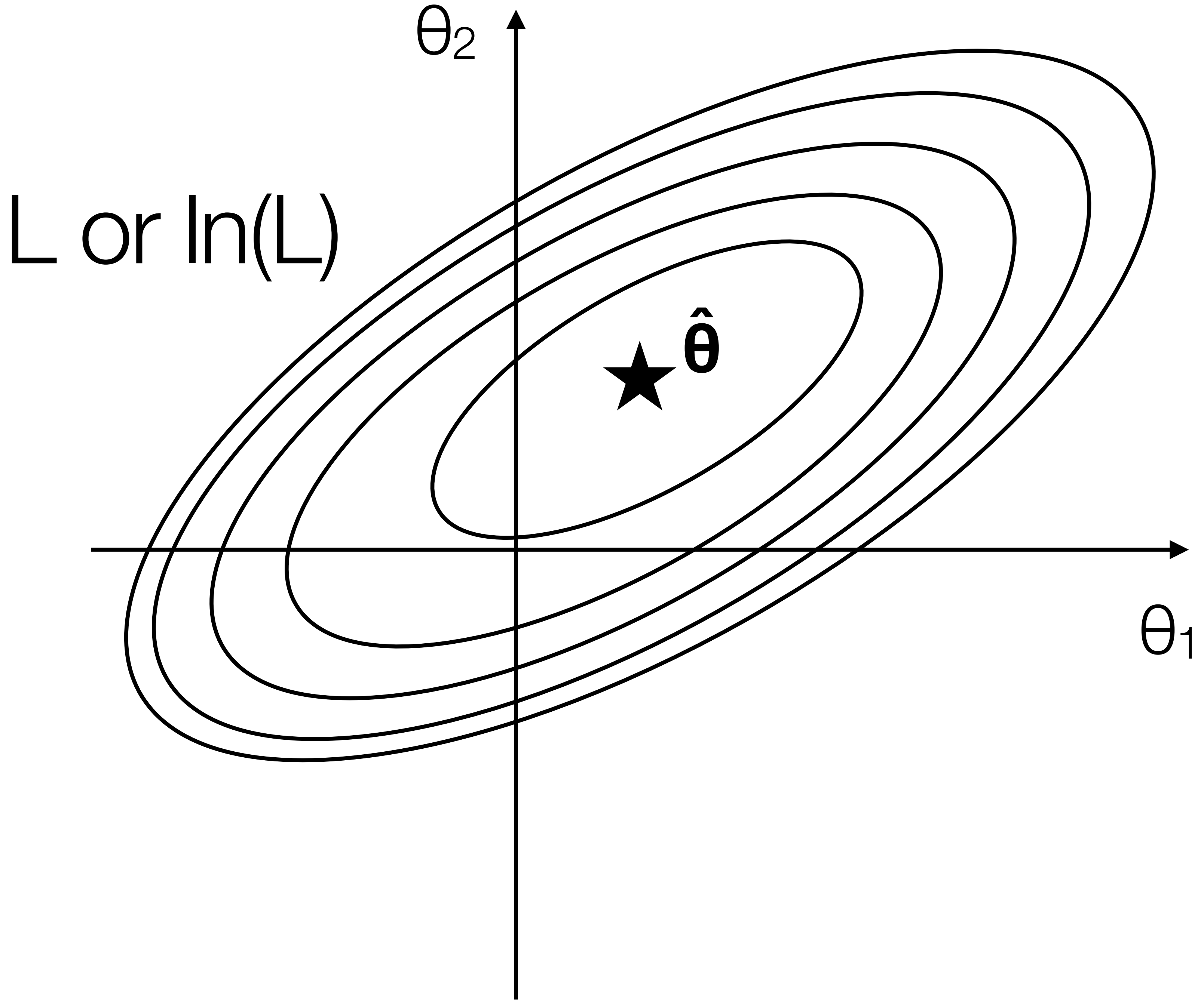
$$\Delta\theta = \left|\frac{\int(\theta-\hat{\theta})^2\mathscr{L}d\theta}{\int\mathscr{L}d\theta}\right|^{1/2}$$

- The procedure for obtaining the maximum likelihood solution is to solve the M simultaneous equations

$$\frac{\partial w}{\partial\theta_k} = 0 \quad \text{where} \quad w = \ln\mathscr{L}(\theta_1,\ldots,\theta_M)$$

Maximum Likelihood (ML) estimators

The most used method for constructing estimators is to take the value of the parameter(s) that maximize the likelihood (or equivalently the log-likelihod):



$$\hat{\theta} = \max_{\theta} L(x\,|\,\theta)$$

Possible ML estimators:

- closed-form function of the data
- (more often) numerically.

Example: exponential pdf (1/2)

- Consider exponential pdf: $f(t\,|\,\tau) = \frac{1}{\tau} e^{-t/\tau}$, with a set of independent data: $t_1, \ldots, t_n$

Then the likelihood function is: $L(\tau) = \prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau}$

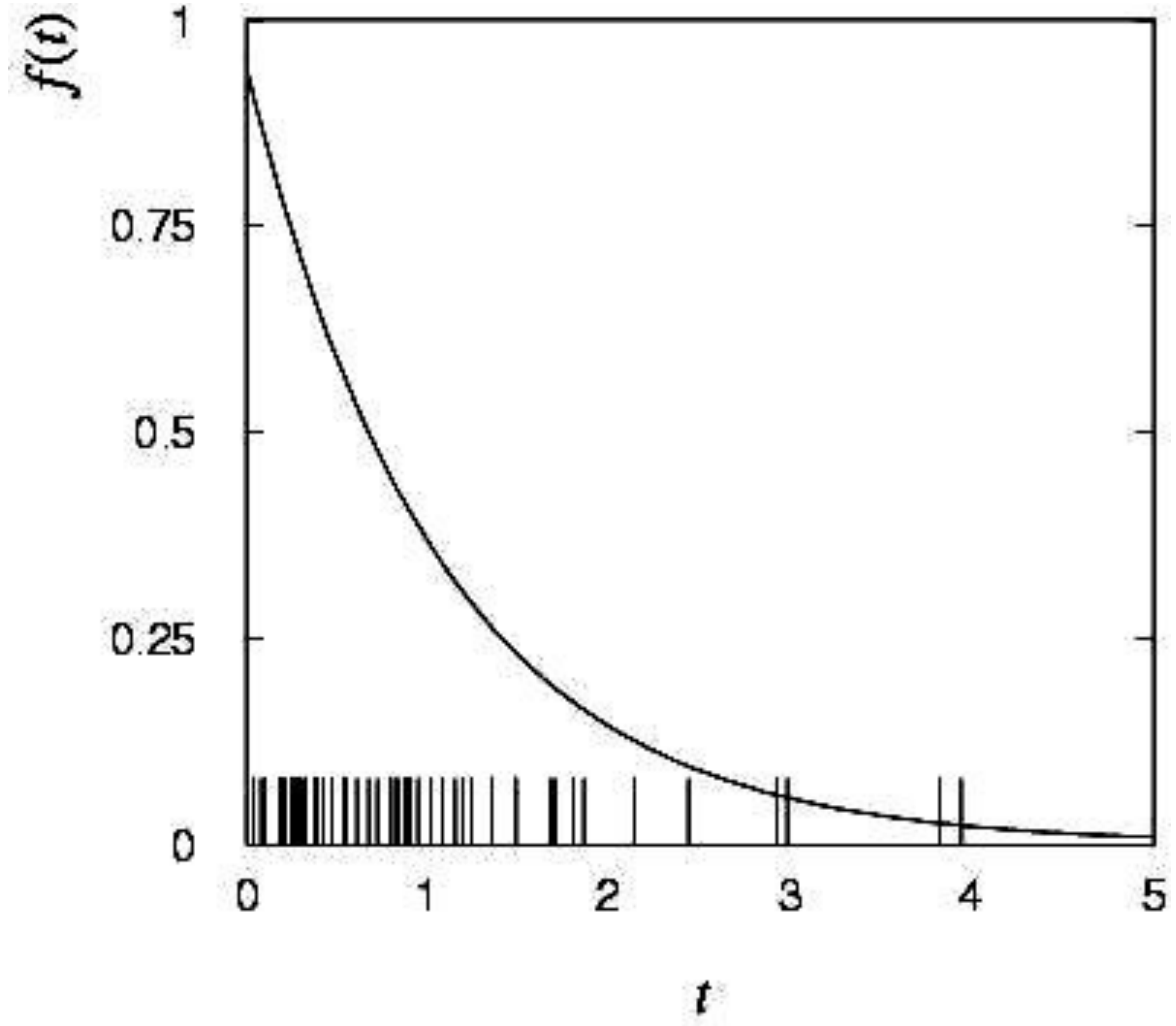
Define the log-likelihood function as the logarithm of L(τ): w≡$\ln L(\tau) = \sum_{i=1}^{n} \ln f(t_i\,|\,\tau) = \sum_{i=1}^{n} \left(\ln \frac{1}{\tau} - \frac{t_i}{\tau} \right)$

L(τ) and lnL(τ) get the maximum at the same τ value

→ Find ML maximum by setting: $\frac{\partial \ln L(\tau)}{\partial \tau} = 0$

$\sum_i \left(-\frac{1}{\tau} + \frac{t_i}{\tau^2} \right) = 0$

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} t_i$$



Example: exponential pdf (2/2)

Now for the mean and variance:

$$E[t] = \int_0^\infty t f(t\,|\,\tau)dt = \int_0^\infty t\frac{1}{\tau}e^{-t/\tau}dt = -\,e^{-t/\tau}(\tau + t)\Big|_0^\infty = \tau$$

$$V[t] = \int_0^\infty (t - E[t])^2 f(t\,|\,\tau)dt = \int_0^\infty (t-\tau)^2\frac{1}{\tau}e^{-t/\tau}dt = -\,e^{-t/\tau}(\tau^2 + t^2)\Big|_0^\infty = \tau^2$$

Considering then the ML estimator $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} t_i$

$$E[\hat{\tau}] = E\left[\frac{1}{n}\sum_{i=1}^{n} t_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[t_i] = \tau \qquad b = E[\hat{\tau}] - \tau = 0$$

$$V[\hat{\tau}] = V\left[\frac{1}{n}\sum_{i=1}^{n} t_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} V[t_i] = \frac{\tau^2}{n} \qquad \sigma_{\hat{\tau}} = \frac{\tau}{\sqrt{n}}$$

… a general result

Example: binned exponential

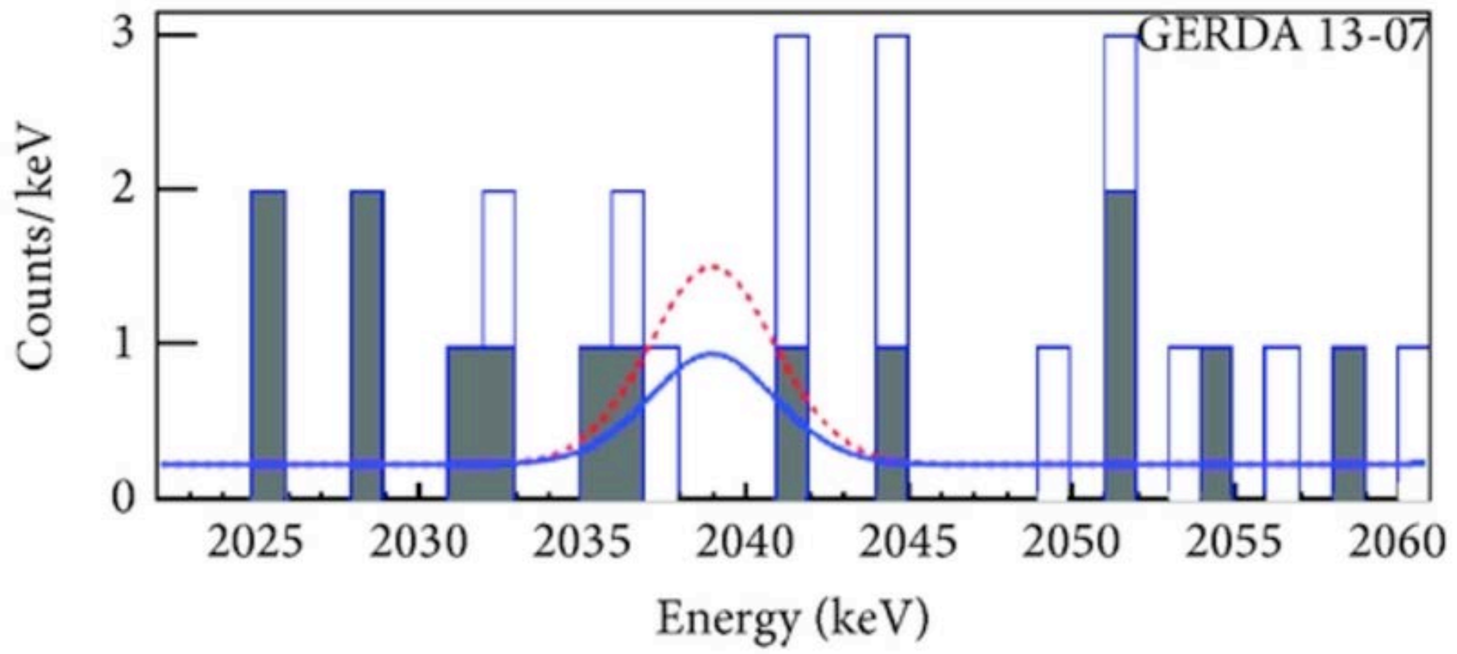
- In the previous example we considered unbinned data: each occurrence time was available separately
- In many cases data are collected in a histogram: n={n$_1$,…,n$_N$} where n$_k$ are the counts integrated over the bin width
- In this case (mainly depending on the number of events) we can use for $f(t\,|\,\tau)$ a multinomial or a Poisson

Let's consider a Poisson distribution: $f(n_k\,|\,\tau) = \mu_k^{n_k} e^{-\mu_k}/n_k!$ with $\mu_k = \int_{t_{k-1}}^{t_k} e^{-t/\tau}/\tau = e^{-t_k/\tau} {-}^{-t_{k-1}/\tau}$

Then $w = \ln L(\tau) = \sum_{i=1}^{N} \ln f(t_i\,|\,\tau) = \sum_k (n_k \ln \mu_k - \mu_k - \ln n_k!)$

- Now we should equate to zero the derivative with respect to τ and equate solve the corresponding equation.
- This is a bit cumbersome but can be easily solved numerically (e.g. finding the minimum of -w)

Note: the term $\sum_k \ln n_k!$ does not depend on τ



Example: Gaussian distribution

Let's consider the measurement of a physical parameter θ₀, where x is known to have a measuring error σ:

$$f(x|\theta_0) = \frac{1}{\sqrt{2\pi\sigma}} \exp[-(x-\theta_0)^2/2\sigma^2]$$

For a set of measurements{x$_k$} each with an error σ$_k$:

$$\mathscr{L}(\theta) = \prod_k \frac{1}{\sqrt{2\pi\sigma_k}} \exp[-(x_k-\theta_0)^2/2\sigma_k^2]$$

Then

$$\boxed{w = -\frac{1}{2}\sum_k \frac{(x_k-\theta)^2}{\sigma_k^2} + C} \quad \chi^2$$

$$\frac{\partial w}{\partial \theta} = \sum_k \frac{x_k - \theta}{\sigma_k^2} \quad \text{or} \quad \sum_k \frac{x_k}{\sigma_k^2} - \sum_k \frac{\hat{\theta}}{\sigma_k^2} = 0$$

The maximum Likelihood (ML) solution is therefore

$$\boxed{\hat{\theta} = \frac{\sum \frac{x_k}{\sigma_k^2}}{\sum \frac{1}{\sigma_k^2}} \quad \text{i.e. the weighted mean}}$$

Note: If σ$_k$=σ $\hat{\theta} = \frac{\sum x_k}{N}$

ML errors: one parameter

It can be shown that asymptotically (i.e. for large numbers) $\mathscr{L}$ approaches a Gaussian:

$$\mathscr{L}(\theta) \propto \exp[-(h/2)(\theta - \hat{\theta})^2]$$

where $1/\sqrt{h}$ is the rms spread of θ about θ̂

Therefore

$$w = -\frac{h}{2}(\theta - \hat{\theta})^2 + C, \qquad \frac{\partial w}{\partial \theta} = h(\theta - \hat{\theta}), \quad \text{and} \quad \frac{\partial^2 w}{\partial \theta^2} = -h$$

and thus

$$\Delta\theta = \left[-\frac{\partial^2 w}{\partial \theta^2}\right]^{-1/2}$$

ML Parabolic error

When combining different results

$$\frac{\partial^2 w}{\partial \theta^2} = \sum \frac{-1}{\sigma_k^2}, \qquad \text{from which} \quad \Delta\theta = \left[\sum \frac{1}{\sigma_k^2}\right]^{-1/2}$$

error combination law

In many actual problems, even though neither θ̂ nor Δθ may be found analytically, a numerical evaluation of $\mathscr{L}$ is always possible.

In case $\mathscr{L}$ is not Gaussian (constant second derivative) we can consider the average error:

$$\overline{\frac{\partial^2 w}{\partial \theta^2}} = \frac{\int (\partial^2 w/\partial \theta^2)\mathscr{L} d\theta}{\int \mathscr{L} d\theta}$$

Variance of estimators

The previous result can be obtained also in a more general way

Information inequality sets a lower bound on the variance of any estimator:

$$V[\hat{\theta}] \geq \left(1 + \frac{\partial b}{\partial \theta}\right)^2 \Big/ E\left[-\frac{\partial^2 \ln L}{\partial \theta^2}\right] \qquad b = E[\hat{\theta}] - \theta$$

(Rao-Cramer-Frechet inequality)

(equality if b=0)

Generally b is small, and equality is a good approximation (e.g. large data sample limit):

$$V[\hat{\theta}] \approx -1 \Big/ E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]$$

“**Parabolic error**”: 2nd derivative of ln L at its maximum:

$$V[\hat{\theta}] = -\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right)^{-1}\Bigg|_{\theta=\hat{\theta}}$$

Variance of estimators (practical method)

Let's expand ln L (θ) about its maximum:

$$\ln L(\theta) = \ln L(\hat{\theta}) + \left[\frac{\partial \ln L}{\partial \theta}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2!}\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + \ldots$$

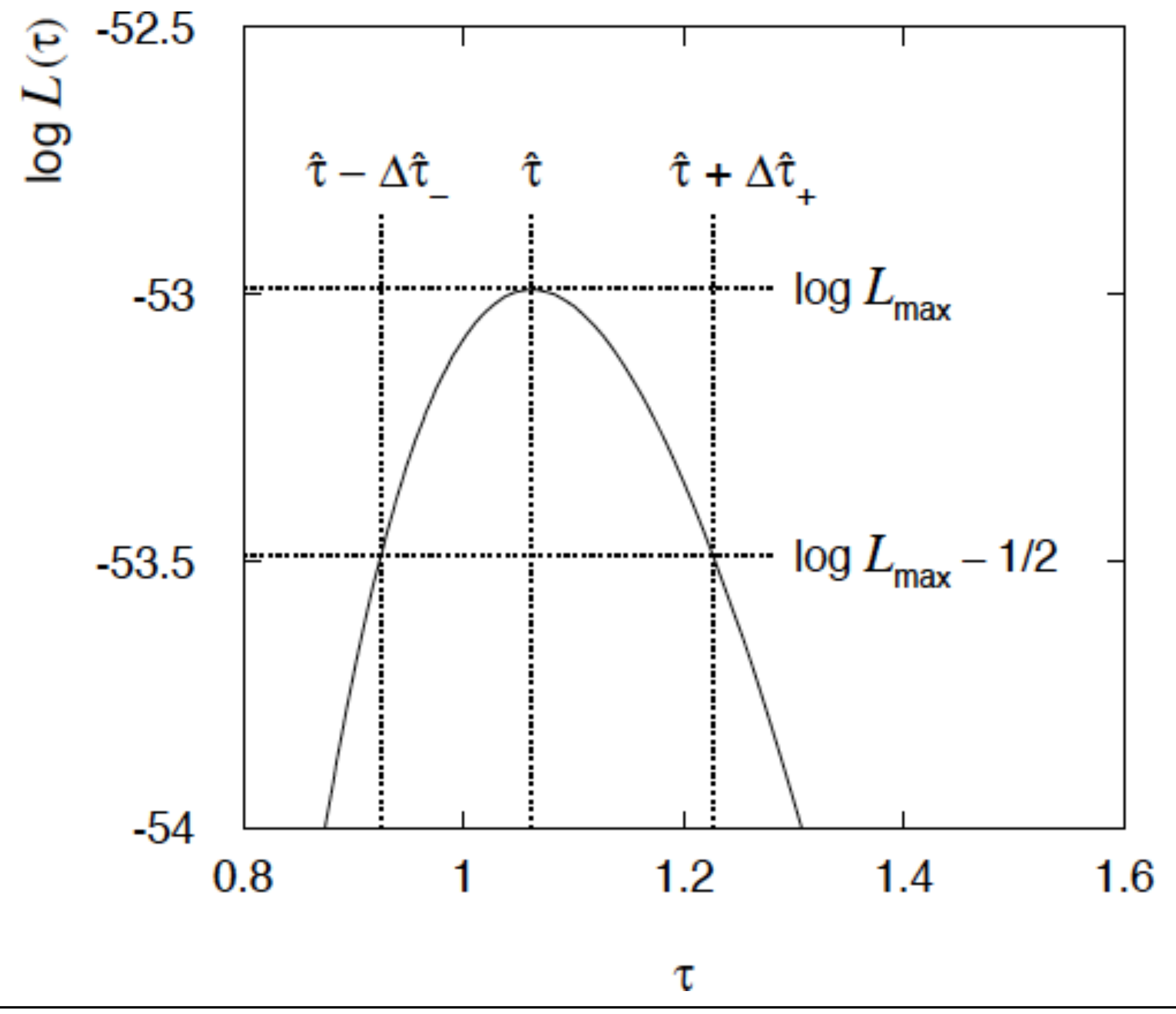
First term is ln L_{max}, second term is zero, for third term use information inequality (~equality):

$$\ln L(\theta) \approx \ln L_{\max} - \frac{(\theta - \hat{\theta})^2}{2\widehat{\sigma^2}_{\hat{\theta}}} \qquad \text{i.e.,} \qquad \ln L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) \approx \ln L_{\max} - \frac{1}{2}$$

Therefore to get $\hat{\sigma}_{\hat{\theta}}$ move θ away from $\hat{\theta}$ until $\ln L$ decreases by 1/2

ML example with exponential:

$$\begin{aligned} \hat{\tau} &= 1.062 \\ \Delta\hat{\tau}_- &= 0.137 \\ \Delta\hat{\tau}_+ &= 0.165 \\ \hat{\sigma}_{\hat{\tau}} &\approx \Delta\hat{\tau}_- \approx \Delta\hat{\tau}_+ \approx 0.15 \end{aligned}$$



Not quite parabolic ln *L* since finite sample size (*n* = 50).

ML: correlated errors (1/2)

Lets' now consider the case of M parameters and a single experiment with N events

The previous results are applicable only in the rare case in which the errors are uncorrelated $\overline{(\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)} = 0 \ \forall i, j$

For the general case let's Taylor-expand w(θ) about θ̂:

$$w(\theta) = w(\hat{\theta}) + \sum_{k=1}^{M} \left(\frac{\partial w}{\partial \theta_k}\bigg|_{\hat{\theta}} \right) \beta_k - \frac{1}{2} \sum_i \sum_j H_{ij}\beta_i\beta_j + \ldots \quad \text{where} \quad \beta_i \equiv \theta_i - \hat{\theta}_i \quad \text{and} \quad H_{ij} \equiv -\frac{\partial^2 w}{\partial\theta_i\partial\theta_j}\bigg|_{\hat{\theta}}$$

Now the second term vanishes ($\partial w/\partial\theta_i|_{\hat{\theta}} = 0$ for all i) and

$$\ln \mathscr{L}(\theta) = w(\hat{\theta}) - \frac{1}{2}\sum_i \sum_j H_{ij}\beta_i\beta_j + \ldots \quad \text{or} \quad \mathscr{L}(\theta) \sim c\exp[-\frac{1}{2}\sum_i\sum_j H_{ij}\beta_i\beta_j] \quad \text{i.e. an M-dim G surface}$$

Then it can be shown (simple transformation) that

$$\mathrm{cov}(\theta_\mathrm{i}, \theta_\mathrm{j}) = \overline{(\theta_\mathrm{i} - \hat{\theta}_\mathrm{i})(\theta_\mathrm{j} - \hat{\theta}_\mathrm{j}} = (\underline{\mathrm{H}}^{-1})_{\mathrm{ij}} \quad \text{again with} \quad H_{ij} = -\frac{\partial^2 w}{\partial\theta_i\partial\theta_j}$$

or equivalently

$$\mathscr{L}(\theta) \sim c\exp[-\frac{1}{2}\beta \cdot \mathbf{V}^{-1} \cdot \beta]$$

Example: multidimensional errors (1/2)

Signals from N monoenergetic electrons are Gaussian-distributed with mean θ_1 and standard deviation θ_2. Find $\hat{\theta}_1$, $\hat{\theta}_2$, and their errors

$$\mathcal{L}(\theta_1, \theta_2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\theta_2} \exp[-(x_i - \theta_1)^2/2\theta_2^2]$$

$$w = -\frac{1}{2}\sum_i \frac{(x_i - \theta_1)^2}{\theta_2^2} - N\ln\theta_2 - \frac{N}{2}\ln(2\pi)$$

$$\frac{\partial w}{\partial \theta_1} = \sum_i \frac{(x_i - \theta_1)}{\theta_2^2}$$

$$\frac{\partial w}{\partial \theta_2} = \frac{1}{\theta_2^3}\sum_i (x_i - \theta_1)^2 - \frac{N}{\theta_2}$$

We get therefore (equating derivatives to zero)

$$\hat{\theta}_1 = \frac{1}{N}\sum_i x_i \quad \text{and} \quad \hat{\theta}_2 = \sqrt{\frac{\sum (x_i - \theta_1)^2}{N}}$$

Someone could notice the difference with respect to the SD formula. Indeed mode ($\hat{\theta}_2$) and mean ($\bar{\theta}_2$) fall in different positions

$$\bar{\theta}_2 = \sqrt{\frac{\sum (x_i - \hat{\theta}_1)^2}{N-1}}$$

Example: multidimensional errors (2/2)

The matrix H is now obtained by evaluating the w derivatives at $\hat{\boldsymbol{\theta}}$

$$\frac{\partial^2 w}{\partial \theta_1^2} = -\frac{N}{\theta_2^2}$$

$$\frac{\partial^2 w}{\partial \theta_2^2} = -\frac{3}{\theta_2^4}\sum (x_i - \theta_1)^2 + \frac{N}{\theta_2^2} = -\frac{2N}{\hat{\theta}_2^2}$$

$$\frac{\partial^2 w}{\partial \theta_1 \partial \theta_2} = -\frac{2}{\theta_2^2}\sum (x_i - \theta_1) = 0$$

Therefore:

$$\underline{H} = \begin{bmatrix} \frac{N}{\hat{\theta}_2^2} & 0 \\ 0 & \frac{2N}{\hat{\theta}_2^2} \end{bmatrix} \quad \text{and} \quad \underline{H}^{-1} = \begin{bmatrix} \frac{\hat{\theta}_2^2}{N} & 0 \\ 0 & \frac{\hat{\theta}_2^2}{2N} \end{bmatrix}$$

The errors on θ$_1$ and θ$_2$ are the square roots of the diagonal elements of H^{-1} :

$$\Delta\theta_1 = \frac{\hat{\theta}_2}{\sqrt{N}} \quad \text{and} \quad \Delta\theta_2 = \frac{\hat{\theta}_2}{\sqrt{2N}}$$

The error on the mean θ$_1$ is $\hat{\theta}_1/\sqrt{N}$ and on the standard deviation θ$_2$ is $\hat{\theta}_2/\sqrt{2N}$

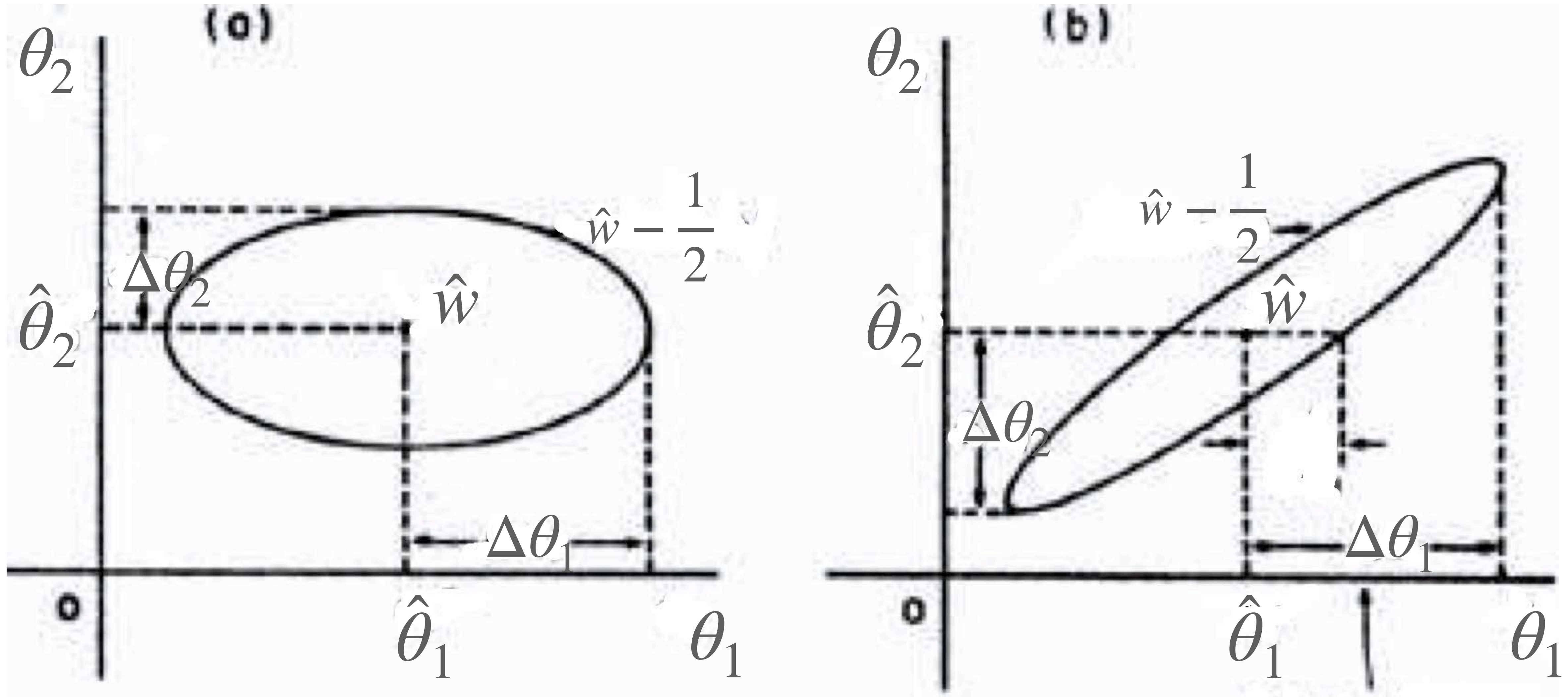
Covariance or error matrix

The covariance matrix $V_{ij} \equiv \overline{(\theta_i - \hat{\theta}_i)(\theta_j - \hat{\theta}_j)}$ can be computed through $\underline{V} = \underline{H}^{-1}$ and $H_{ij} = -\dfrac{\partial^2 w}{\partial\theta_i \partial\theta_j}$

The diagonal elements of V̲ are the variances of the θ's.
If V_{ij}=0 ∀i≠j, errors are uncorrelated (contour plots of w are ellipses whose major axes are the errors on θ_j):
$\Delta\theta_j = [H_{jj}]^{-1}$.

Contours of constant w as a function of θ1, and θ2:

- Maximum likelihood solution is at w = ŵ.
- Errors in θ1, and θ2 are obtained from ellipse where w = (ŵ - 1/2
  - (a) Uncorrelated errors
  - (b) Correlated error
- In either case $\Delta\theta_1^2 = V_{11} = (H^{-1})_{11}$ and $\Delta\theta_2^2 = V_{22} = (H^{-1})_{22}$



- In In cases where H̲ cannot be evaluated analytically, the θ̂'s can be found numerically and the errors can be found by Plotting the ellipsoid where w is 1/2 unit less than ŵ.
- The extremums of this ellipsoid are the rms error in the ŵ's.
- One should allow all the θ_j to change freely and search for the maximum change in θ_j which makes w = (ŵ - 1/2).
- This maximum change in θ_j, is the error in θ_j and is $\sqrt{V_{jj}}$.

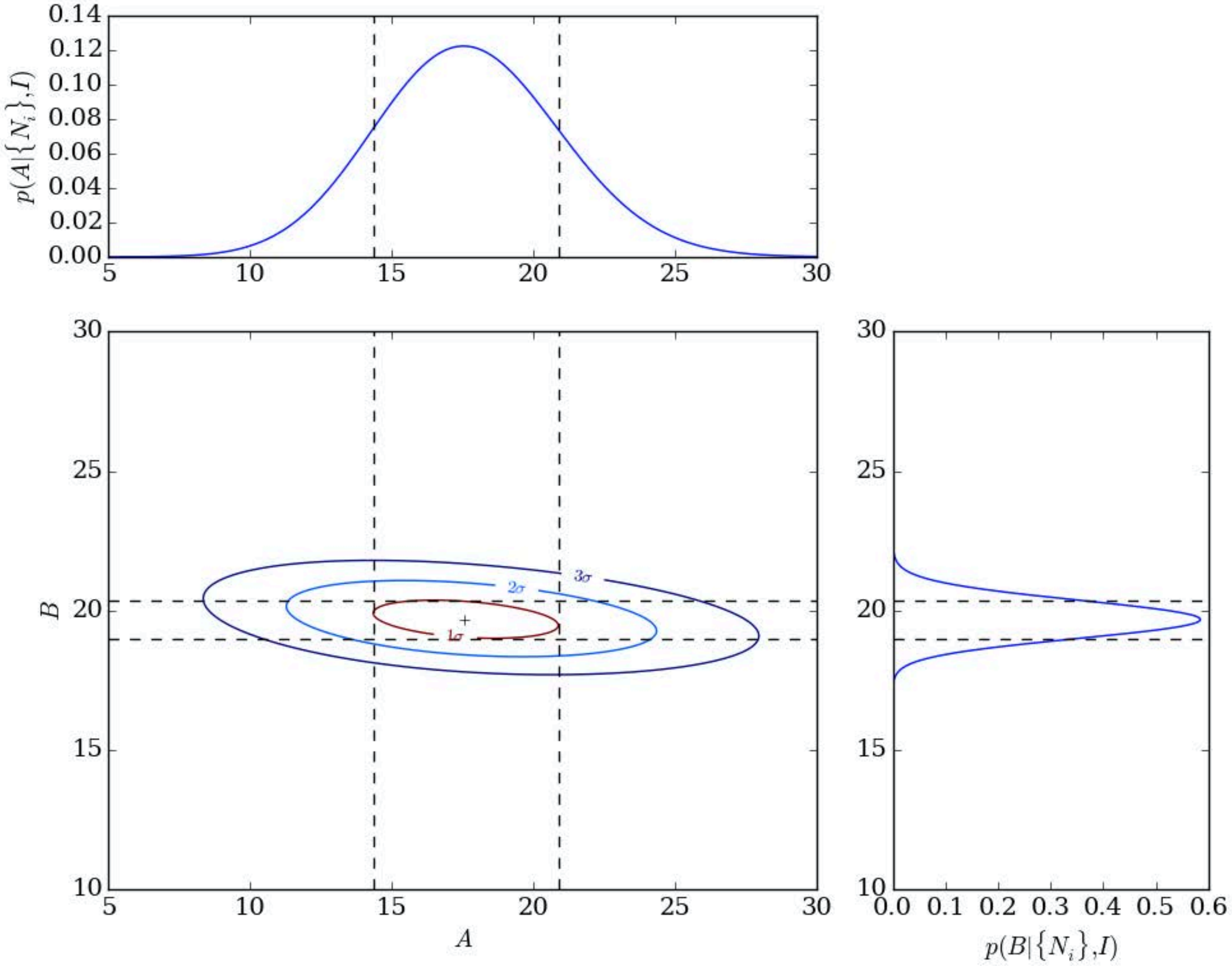
➡ MINUIT: MINOS errors

ML variances: >1 parameter

- For > 1 parameters:

$$\mathrm{cov}(\theta_\mathrm{i}, \theta_\mathrm{j}) = \left(- \frac{\partial^2 \ln \mathscr{L}}{\partial\theta_\mathrm{i} \partial\theta_\mathrm{j}} \bigg|_{\hat{\theta}_\mathrm{i}\hat{\theta}_\mathrm{j}} \right)^{-1}$$

- Use the ΔlnL trick to get contours for 1σ, 2σ, etc.
- Project ellipse onto each axis (i.e., marginalize) to get uncertainties in each parameter



Multiparametric case

- Usually we calculate a joint likelihood on several parameters but only produce confidence intervals for individual parameters.
- However, if we want confidence ellipses in several parameters jointly, we need to change the $\Delta\ln\mathscr{L}$ rule a bit

$\Delta\ln\mathscr{L}$ values as a function of probability for n joint parameters

		joint parameters					
Range	**p**	n=1	n=2	n=3	n=4	n=5	n=6
1σ	**68.3**	0.50	1.11	1.76	2.36	2.95	3.52
2σ	**95.4**	2.00	3.09	4.01	4.85	5.65	6.40
3σ	**99.7**	4.50	5.90	7.10	8.15	9.10	10.05

However usually we are interested in the marginal distributions of individual parameters

Propagation of errors

Suppose we have a single function of our parameters: $y=y(\boldsymbol{\theta})$. Then $\hat{y}=y(\hat{\boldsymbol{\theta}})$.

To find the errors on y:

$$y-\hat{y}=\sum\frac{\partial y}{\partial\theta_k}(\theta_k-\hat{\theta}_k)$$

$$\overline{(y-\hat{y})^2}=\sum_j\sum_k\frac{\partial y}{\partial\theta_j}\frac{\partial y}{\partial\theta_k}\overline{(\theta_j-\hat{\theta}_j)(\theta_k-\hat{\theta}_k)}$$

$$(\Delta y)_{rms}=\sqrt{\sum_j\sum_k\frac{\partial y}{\partial\theta_j}\frac{\partial y}{\partial\theta_k}V_{jk}}$$

Then, if the variables are uncorrelated:

$$(\Delta y)_{rms}=\sqrt{\sum_j\left(\frac{\partial y}{\partial\theta_j}\right)^2(\Delta\theta_j)^2}$$

which is the well known formula of error propagation

If instead $\mathbf{y}(\boldsymbol{\theta})$ is a one to one correspondence (basis change) and the θ_k error matrix is known, then

$$\overline{(y_i-\hat{y}_i)(y_j-\hat{y}_j)}=\sum_m\sum_n\frac{\partial y_i}{\partial\theta_m}\frac{\partial y_i}{\partial\theta_n}H^{-1}_{mn}$$

In practice $\frac{\partial y_i}{\partial\theta_m}$ is not always calculable but

$\frac{\partial\theta_i}{\partial y_m}=J_{ij}$ is generally easy to compute and

$$\frac{\partial y_i}{\partial\theta_m}=(\underline{J}^{-1})_{im}$$

Example: correlated errors (scalar)

If f is a scalar function of the model parameters, the previous result can be rewritten as

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + 2\rho \left(\frac{\partial f}{\partial x}\right)\left(\frac{\partial f}{\partial y}\right)\sigma_x\sigma_y \quad \text{with } \rho = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}}$$

where f is some function f(x,y) of the data

Example:

- Collect N$_T$ events: N$_F$ forward, N$_B$ backward
- Evaluate error on R=N$_F$/N$_T$

Assume:

- Everything Poisson
- F and B uncorrelated
- F and T correlated

$$\langle N_F(N_F+N_B)\rangle - \langle N_F\rangle\langle N_F+N_B\rangle = \langle N_F^2\rangle + \langle N_F N_B\rangle - \langle N_F\rangle\langle N_F\rangle - \langle N_F\rangle\langle N_B\rangle$$

solution:

- $cov(F,T) = \langle N_F N_T\rangle - \langle N_F\rangle\langle N_T\rangle = V(N_f) = N_F$
  since $\langle N_F N_B\rangle = \langle N_F\rangle\langle N_B\rangle$
- Then $\rho \equiv \dfrac{\mathrm{cov(N_F, N_T)}}{\sigma_F\sigma_T} = \dfrac{N_F}{\sqrt{N_F N_T}} = \sqrt{\dfrac{N_F}{N_T}}$
- $\sigma_R^2 = \left(\dfrac{1}{N_T}\right)^2 N_F + \left(\dfrac{-N_F}{N_T^2}\right)^2 N_T + 2\sqrt{\dfrac{N_F}{N_T}}\left(\dfrac{1}{N_T}\right)\left(\dfrac{-N_F}{N_T^2}\right)\sqrt{N_F N_T}$

$$= \frac{N_F N_T + N_F^2 - 2N_F^2}{N_T^3} = \frac{R(1-R)}{N_T}$$

Using R=N$_F$/(N$_F$+N$_B$):

- $\sigma_R^2 = \left(\dfrac{1}{N_T} - \dfrac{N_F}{N_T^2}\right) N_F + \left(\dfrac{-N_F}{N_T^2}\right)^2 N_B = \left(\dfrac{N_B}{N_T^2}\right)^2 N_F + \left(\dfrac{N_F}{N_T^2}\right)^2 N_B = \dfrac{N_F N_B}{N_T^3} = \dfrac{R(1-R)}{N_T}$

Example: correlated errors (vector)

Suppose we want to use radius and acceleration to specify the circular orbit of an electron in a uniform magnetic field; i.e., $y_1 = r$ and $y_2 = a$. Suppose the original measured quantities are $\theta_1 = \tau = (10 \pm 1)$ μs and $\theta_2 = v = (100 \pm 2)$ km/s. Also since the velocity measurement depended on the time measurement, there was a correlated error $\langle \Delta t \Delta v \rangle = 1.5 \cdot 10$ m. Find r, a and their errors.

From the data:

$y_1 = r = v\tau/2\pi = 0.159$ m

$y_2 = a = v^2/r = 2\pi v/\tau = 6.28 \cdot 10^{10}$ m/s^2

Therefore

$y_1 = \theta_1\theta_2/2\pi$

$y_2 = 2\pi\theta_2/\theta_1$

and

$$\frac{\partial y_1}{\partial \theta_1} = \frac{\theta_2}{2\pi}$$

$$\frac{\partial y_1}{\partial \theta_2} = \frac{\theta_1}{2\pi}$$

$$\frac{\partial y_2}{\partial \theta_1} = -\frac{2\pi\theta_2}{\theta_1^2}$$

$$\frac{\partial y_2}{\partial \theta_2} = \frac{2\pi}{\theta_1}$$

The measurement errors specify the error matrix as

$$V = \begin{pmatrix} 10^{12}\,\mathrm{s} & 1.5 \cdot 10^{-3}\,\mathrm{m} \\ 1.5 \cdot 10^{-3}\,\mathrm{m} & 4 \cdot 10^{6}\,\mathrm{m}^2/\mathrm{s}^2 \end{pmatrix}$$

We get therefore:

$$(\Delta y_1)^2 = \left[\frac{\theta_2}{2\pi}\right]^2 V_{11} + 2\left[\frac{\theta_2}{2\pi}\right]\left[\frac{\theta_1}{2\pi}\right] V_{12} + \left[\frac{\theta_1}{2\pi}\right]^2 V_{22} = \frac{v^2}{4\pi^2}V_{11} + \frac{v\tau}{2\pi^2}V_{12}\frac{\tau^2}{4\pi^2}V_{22} = 3.39 \cdot 10^{-4}\ \mathrm{m}^2$$

$$(\Delta y_2)^2 = \left[-\frac{2\pi\theta_2}{\theta_1^2}\right]^2 V_{11} + 2\left[-\frac{2\pi\theta_2}{\theta_1^2}\right]\left[\frac{2\pi}{\theta_1}\right] V_{12} + \left[\frac{2\pi}{\theta_1}\right]^2 V_{22} = 2.92 \cdot 10^{19}\ \mathrm{m}^2/\mathrm{s}^4$$

And the result is:

- r=(0.159 ± 0.018) m
- a = (6.28 ± 0.54)10^{10}m/s^2

Errors from Likelihood: summary

Estimate a model parameter θ̂ by maximizing the likelihood

In the **large N limit**:

i) This is unbiassed and efficient

ii) The error is given by $\frac{1}{\sigma^2} = -\left\langle \frac{d^2 \ln L}{d\theta^2} \right\rangle$

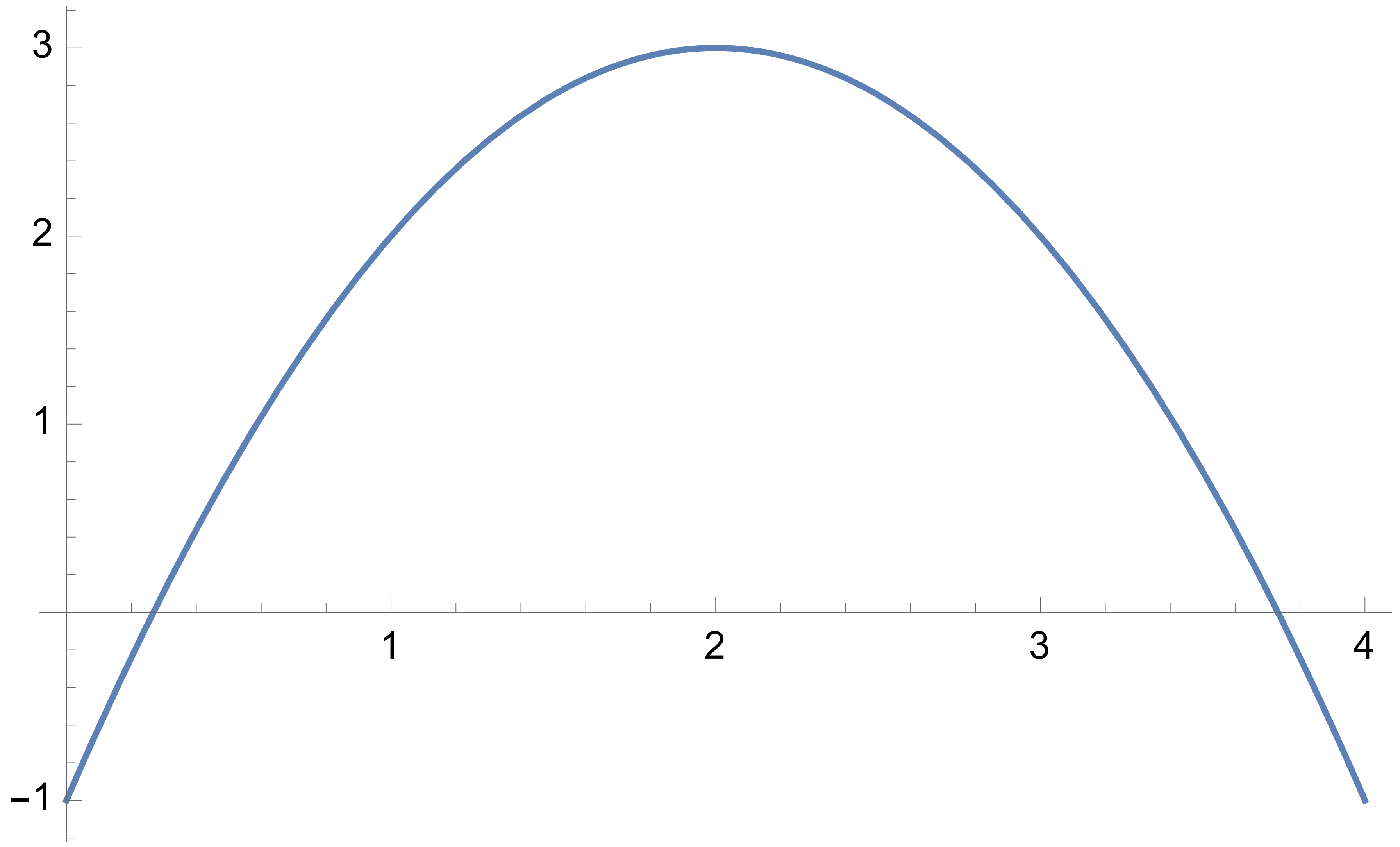
iii) ln L is a parabola: $L = L_{max} - \frac{1}{2}C(\theta - \hat{\theta})^2$

iv) We can approximate $C \equiv -\left.\frac{d^2 \ln L}{d\theta^2}\right|_{\hat{\theta}} = -\left\langle \frac{d^2 \ln L}{d\theta^2} \right\rangle$

v) Read off σ from $\Delta \ln L = -\frac{1}{2}$

vi) Get neats confidence intervals :

- $\Delta \ln L = -1/2$ → 68% CL (1σ, 1 parameter)
- $\Delta \ln L = -2$ → 95.4% CL (2σ, 1 parameter)
- whatever you choose, 2-sided, 1-sided, ….

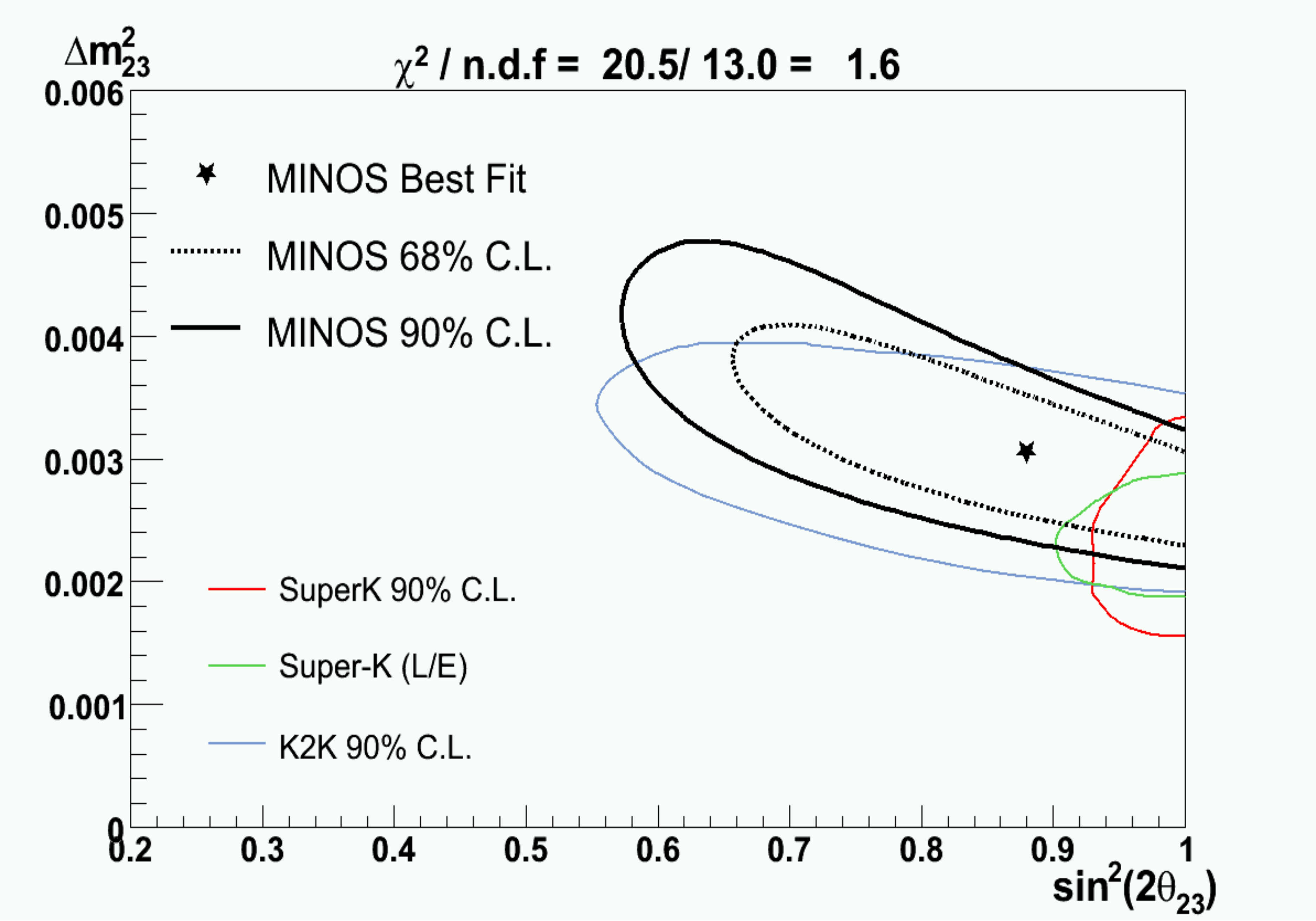


Small N:

- lnL is not a parabola (e.g. asymmetric)
- None of the above is tue
- However:
  - We can transform from θ → θ' parabolic, find the limits, and transform back
  - Would give ΔlnL=-1/2 for 68% CL etc as before
  - Hence asymmetric errors

Small N non-Gaussian measurements

- No longer ellipses/ellipsoids
- Use ΔlnL to define confidence regions, mapping out contours
- Probably not totally accurate, but universal



Alternative: toy Monte Carlo:

- Have dataset
- Take point θ in parameter space
- Test if it is in or out of the 68% (or ...) contour
  - Find $T = \ln L(R\,|\,\hat{\theta}) - \ln(R\,|\,\theta)$ (clearly small T is 'good')
  - Generate many MC sets of R, using θ
  - How often is $T_{MC} > T_{data}$?
  - If more than 68%, M is in the contour

We are ordering the points by their value of T (the Likelihood Ratio) – almost contours but not quite

Hypothesis test

Interval estimation

- Estimation of the parameter is not sufficient.
- It is necessary to analyse and see how **confident** we can be about this particular estimation.
- One way of doing it is defining confidence intervals.
- If we have estimated θ we want to know if the “true” parameter is close to our estimate.
- In other words we want to find an interval that satisfies the following relation:

the probability that the “true” value of parameter θ is in the interval (G_L,G_U) is greater than 1-α

- The actual realisation of the interval (G_L, G_U) is called a (1 - α) confidence interval (usually expressed in %)
  - bounds of the interval are called lower and upper confidence limits
  - (1 - α) is called confidence level

Example:

If population variance is known (σ^2) and we estimate population mean $\overline{x}$, then we expect $Z = \dfrac{\overline{x} - \mu}{\sigma/\sqrt{N}}$ to be normally distributed about the “true” value: N(0,1)

We can find from the table that the probability of Z is more than 1 (less than -1) is equal to 0.1587

Therefore $P(-1 < Z < 1) = P(Z < 1) - P(Z < -1) = 1 - 2 \cdot 0.1587 = 0.6826$ and we conclude

$$P(-1 < \frac{\overline{x} - \mu}{\sigma/\sqrt{N}} < 1) = P(\overline{x} - \sigma/\sqrt{N} < \mu < \overline{x} + \sigma/\sqrt{N}) = 0.6826$$

Conclusion: the confidence level that “true” value is within 1 standard error (standard deviation of sampling distribution) from the sample mean is 0.6826.

Frequentist vs Bayesian

- In the previous example we have transformed a statement about the probability of $\bar{x}$ (data)

  $\bar{x}$ is normally distributed around μ

  which simply tells you that if you repeat the measurement many times you can predict how outcomes will be distributed …

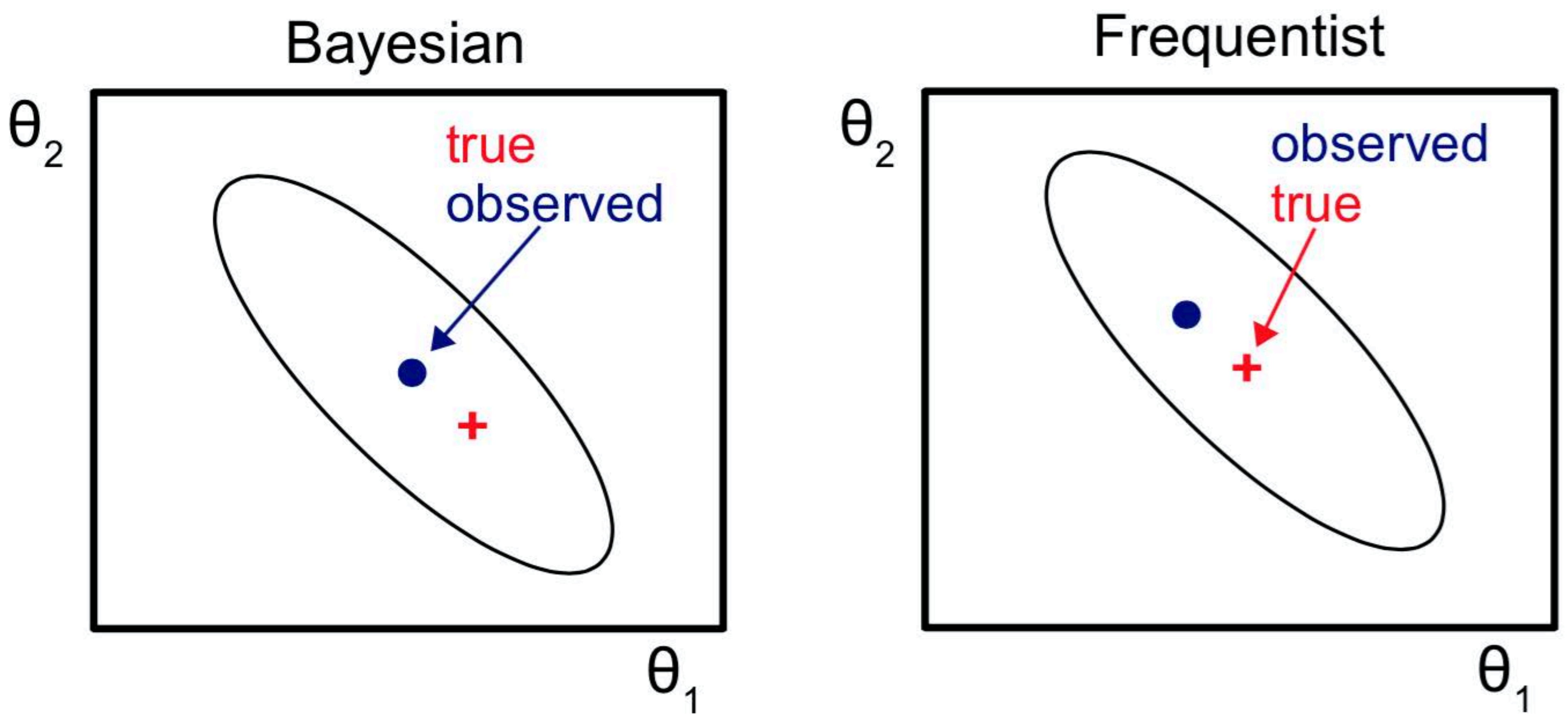
- into a statement about μ: confidence level (probability) that μ lies in the interval $\bar{x} \pm \sigma/\sqrt{N}$ is 68%

Correct interpretation (personal view)

Bayesian: given a measurement, we have some confidence that our best estimate of a parameter lies within some range of the data

Frequentist: given the true value of the parameters, we have some confidence that our measurement lies within some range of the true value

Difference: p(θ|D,I) versus p(D|θ,I)

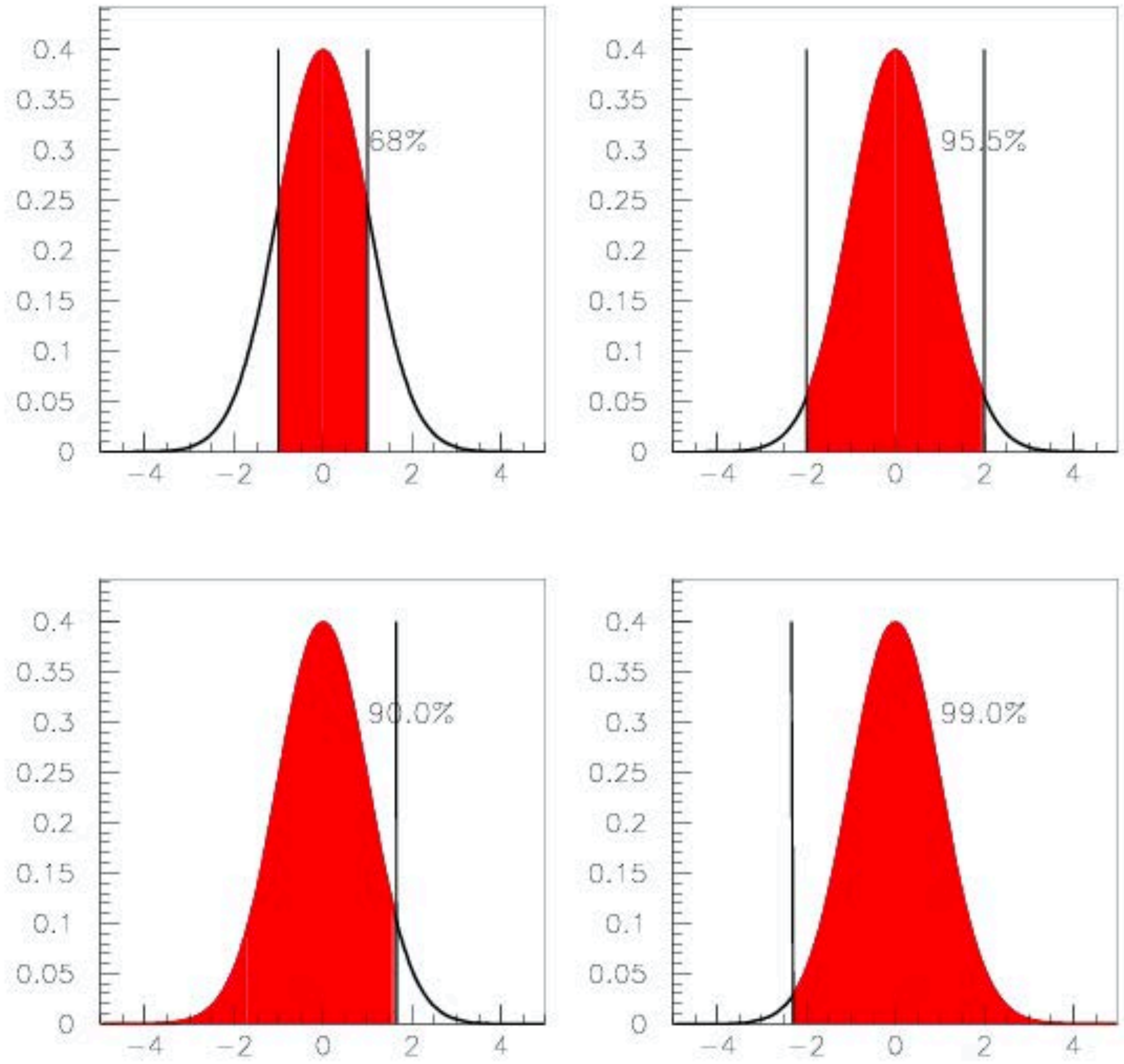


Confidence intervals arbitrariness

For a given Confidence Level …

Some of the possibile choices:

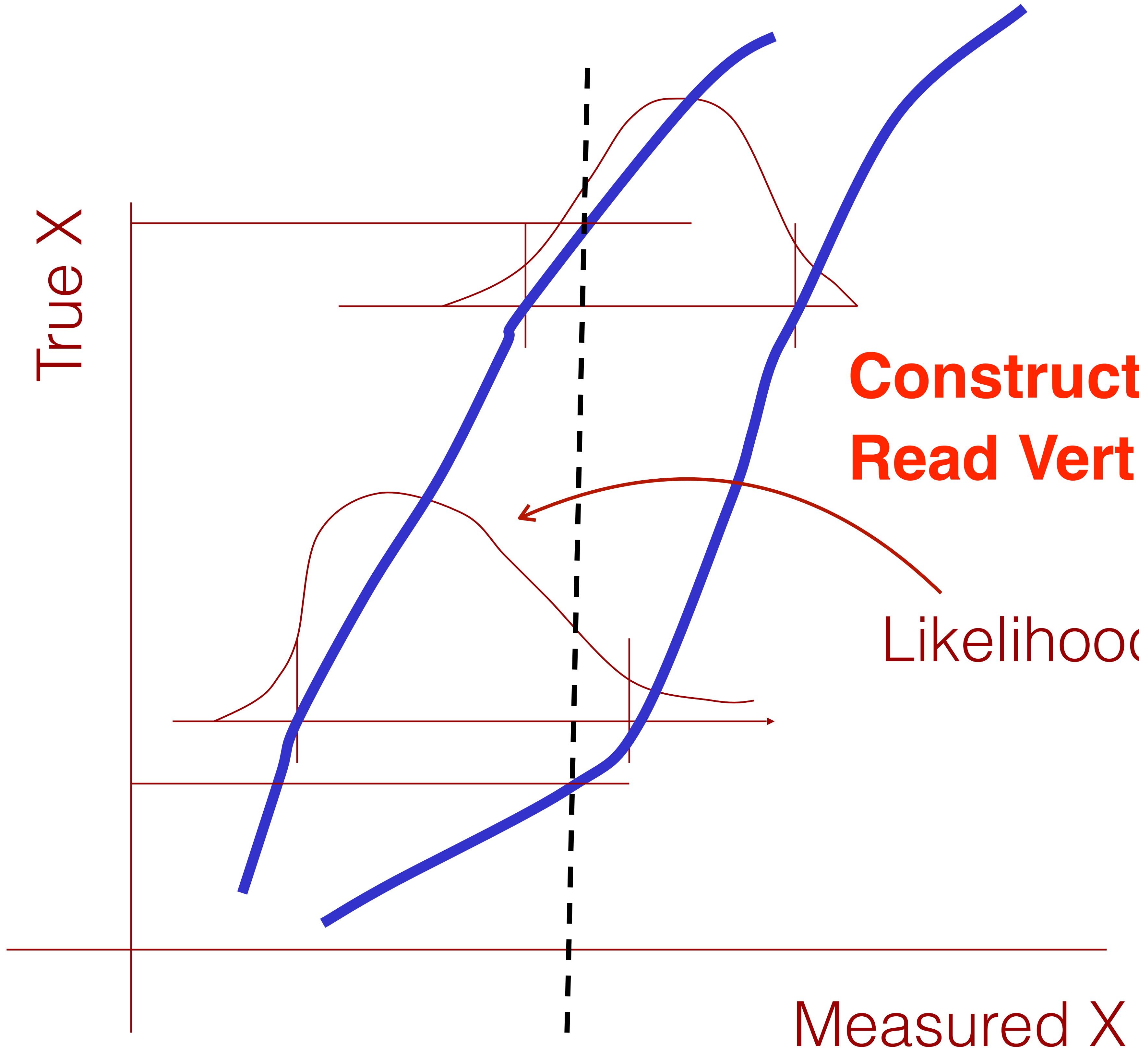
- Upper limit
- Lower limit
- 2-sided limit
  - central
  - shortest
  - …



Inverse probability: confidence belts

Counting experiment:

- For complicated distributions it isn't quite so easy
- But the principle is the same frequentist approach



Construct Horizontally
Read Vertically

Bayes (standard Gaussian measurement)

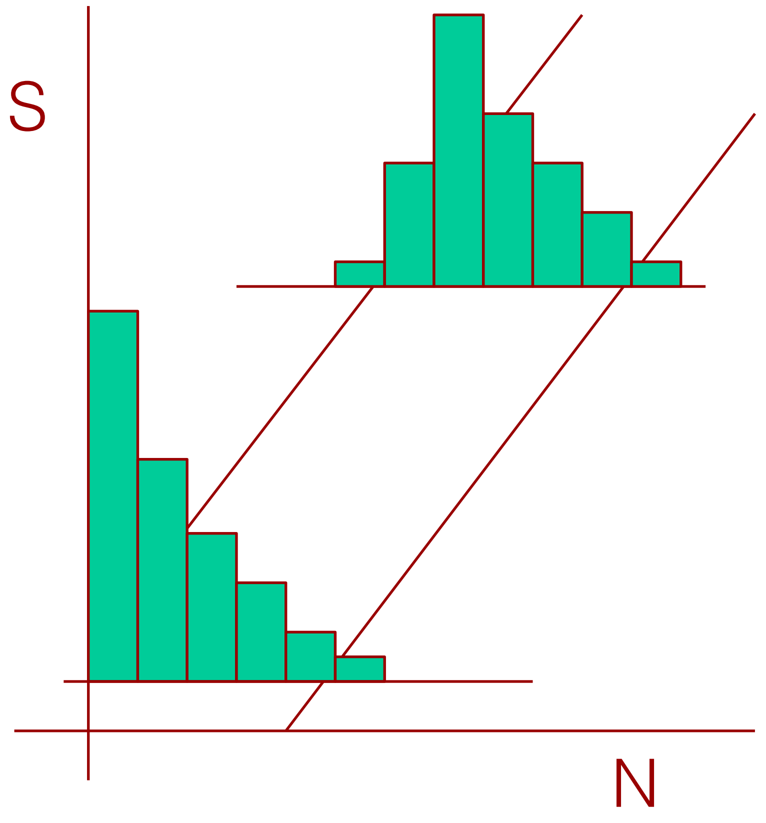
$$P(\text{theory}\,|\,\text{data}) = \frac{\text{P(data}\,|\,\text{theory)P(theory)}}{\text{P(data)}}$$

- No prior knowledge of true value (theory)
- No prior knowledge of measurement result (data)
- P(data|theory) is Gaussian
- P(theory|data) is Gaussian
- Gives same limits as Frequentist method for simple Gaussian
- Interpret this with Probability statements as you prefer

Discrete distributions:

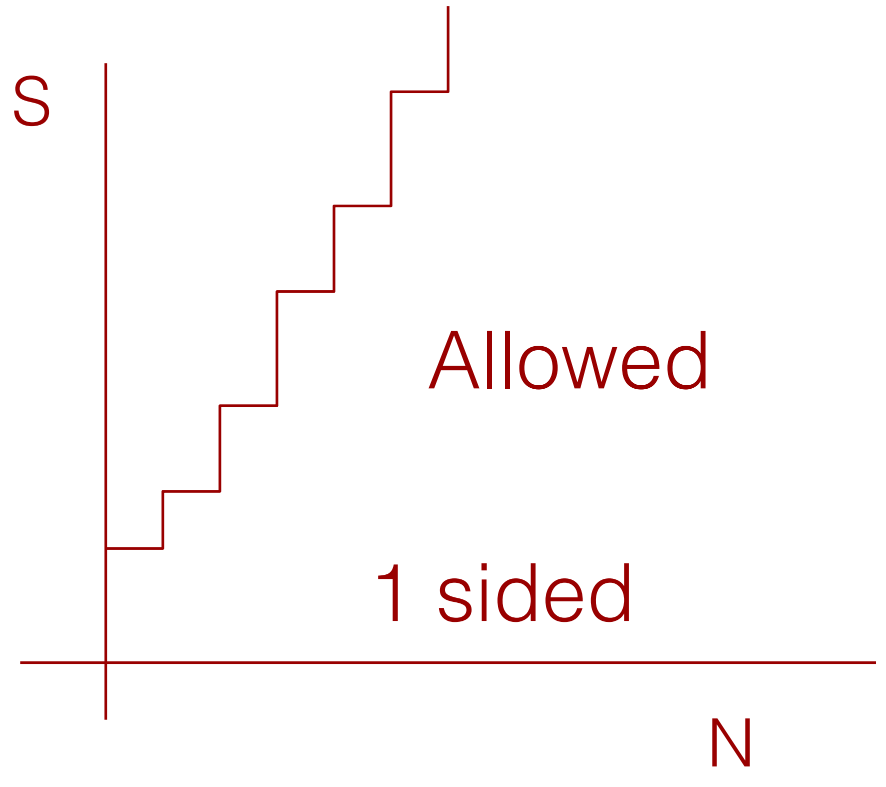
- May be unable to select (say) 5% region
- Play safe.
- Gives overcoverage

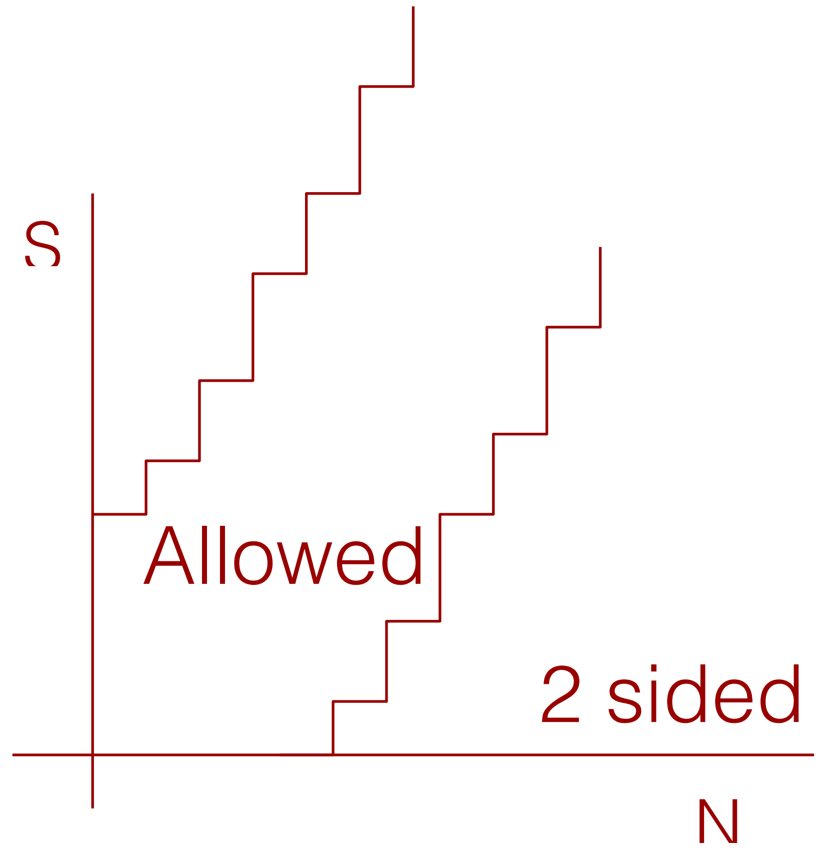
Confidence belts: Feldman-Cousins



Method (90% CL):

- For every S, select N-values in belt
- Total probability must sum to 90%
- **Many possible strategies**

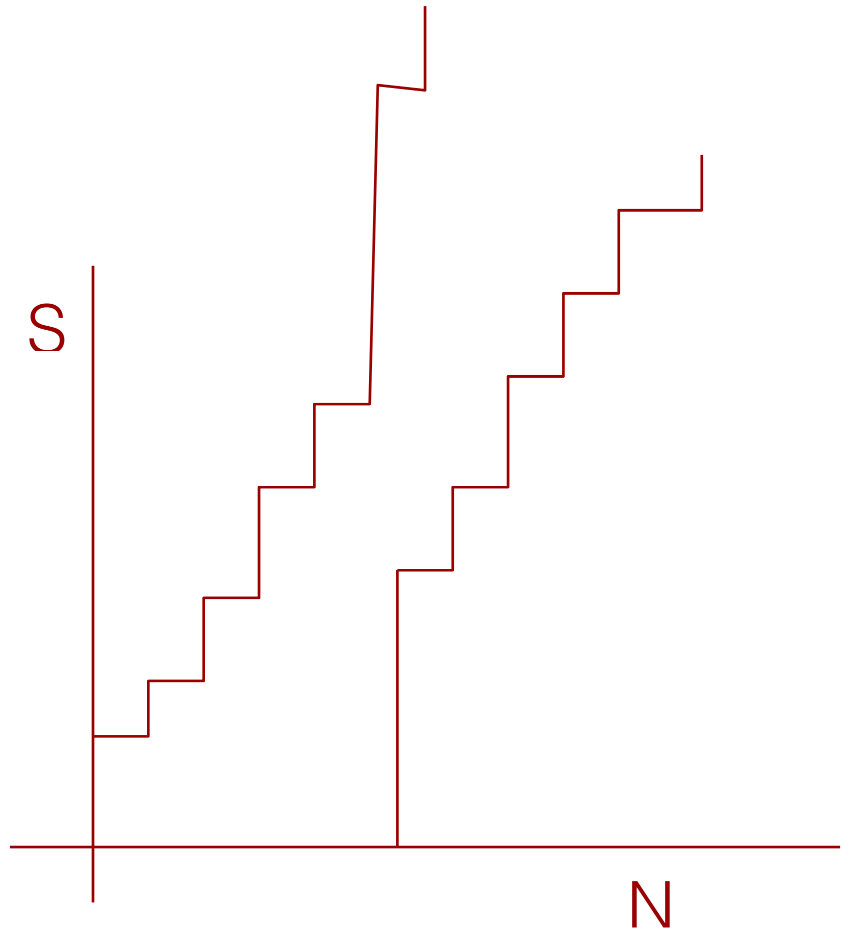




Crow & Gardner:

- Select N-values with highest probability
  → shortest interval

This is not a true confidence belt! Coverage varies.



Feldman & Cousins

- For any given S
- For each N find
  - P(N;S+B)
  - P_{best}=P(N;N) if (N>B) else P(N;B)
- Rank on P/P_{best}
- Accept N into band until **S** P(N;S+B) 90%

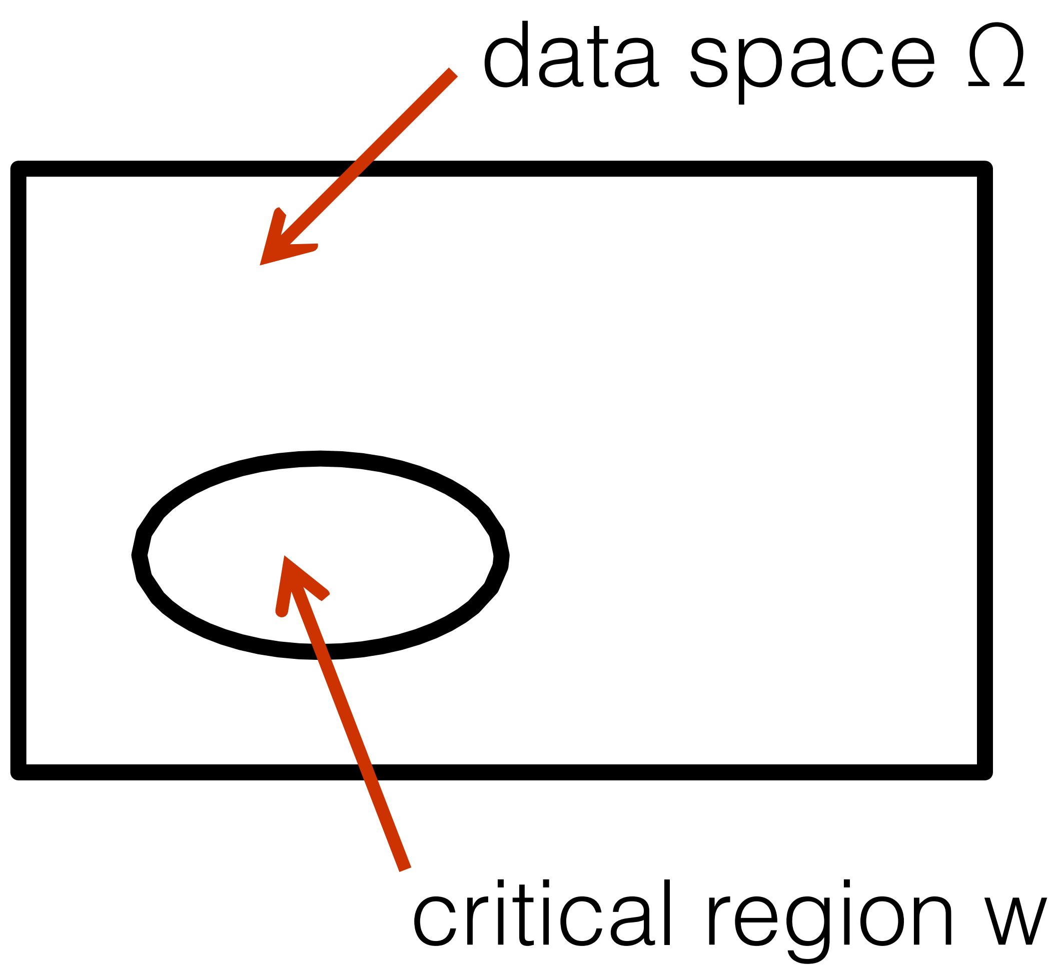
PRO:

- Makes us more honest (a bit)
- Avoids forbidden regions in a Frequentist way

CON:

- Not easy to calculate
- Has to be done separately for each value of B
- Can lead to 2-tailed limits where you don't want to claim a discovery
- Weird effects for N=0; larger B gives lower (=better) upper limit

Frequentist statistical tests



Suppose a measurement produces data x
Consider a hypothesis H_0 ; we want to test an alternative H_1

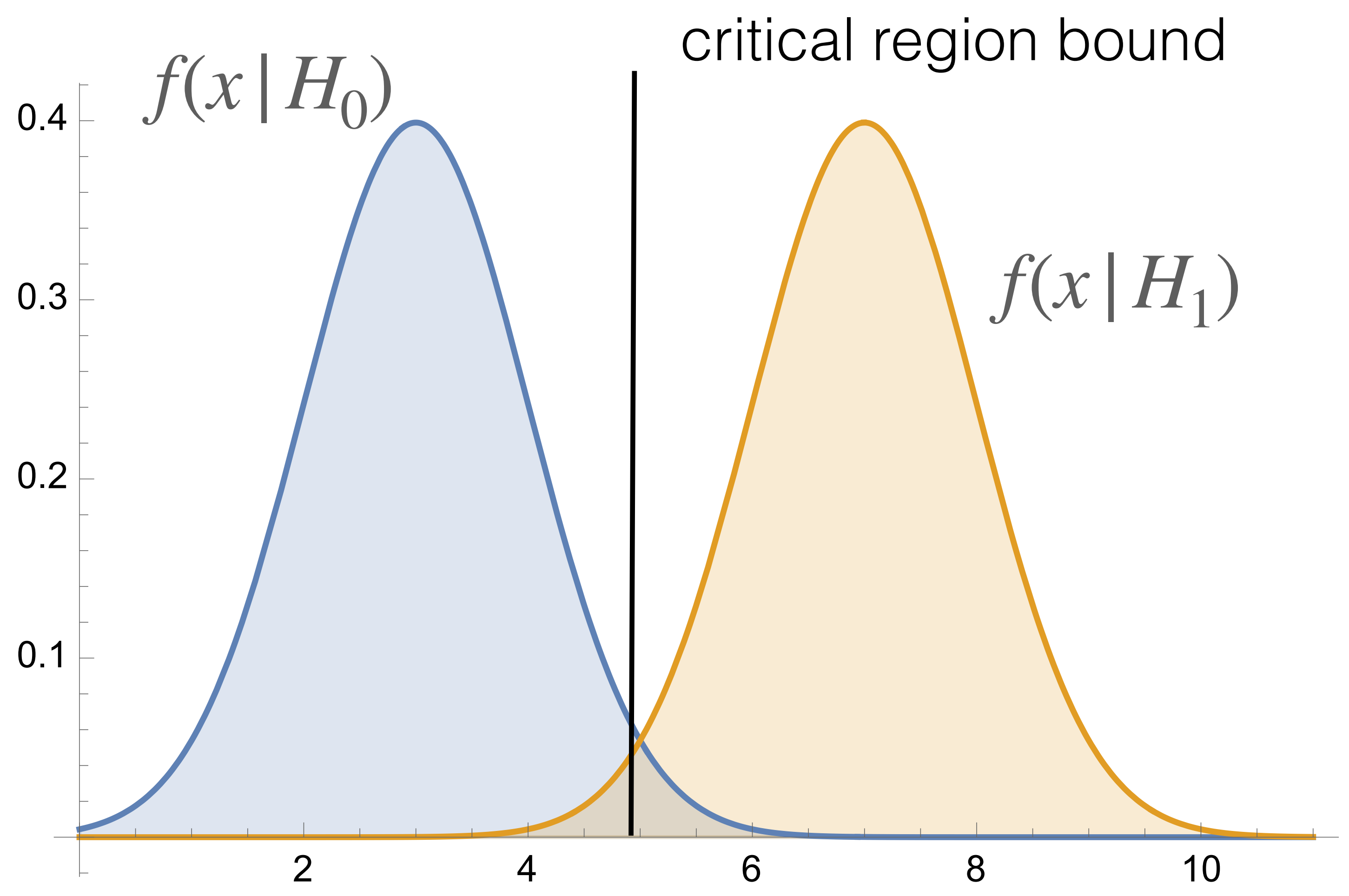
H_0, H_1 specify probability for x: $P(x|H_0)$, $P(x|H_1)$

A test of H_0 is defined by specifying a critical region w of the data space such that there is no more than some (small) probability α, assuming H_0 is correct, to observe the data there

In general: infinite number of possible choices for critical regions with the same significance α.

So the choice of the critical region for a test of H_0 needs to take into account the alternative hypothesis H_1

Place the critical region where there is a low probability for H_0 true, but high for H_1 true



ML: confidence intervals arbitrariness

- While the ML solution does not depend on the choice of the variables (given two different variables λ and τ, then $\frac{\partial w}{\partial \lambda} = \frac{\partial w}{\partial \tau}\frac{\partial \tau}{\partial \lambda} = 0$) the same is **not true for confidence intervals**.

- Confidence intervals are defined by

$$P(\theta' < \theta < \theta'') = \int_{\theta'}^{\theta''} \mathscr{L} d\theta \Big/ \int_{-\infty}^{+\infty} \mathscr{L} d\theta$$

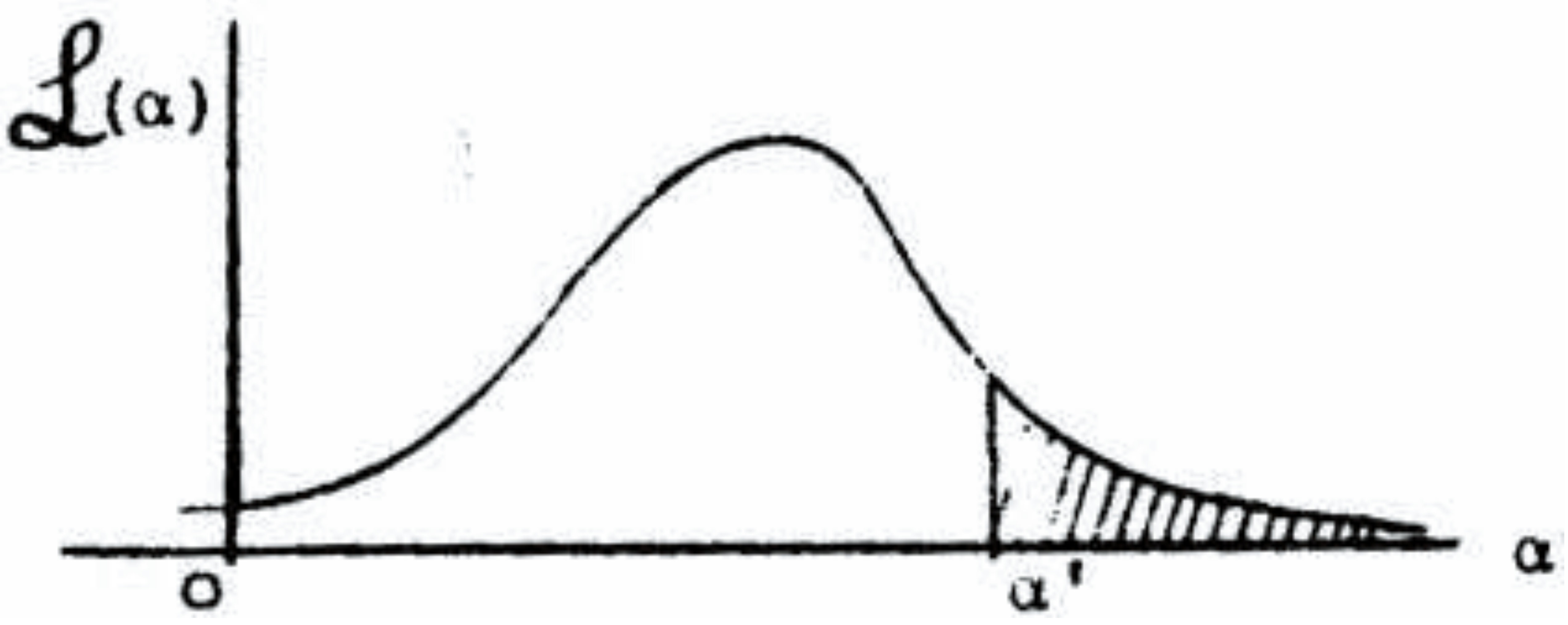
and if we consider the confidence limit:

$$P(\theta > \theta') = \int_{\theta'}^{\infty} \mathscr{L} d\theta \Big/ \int_{-\infty}^{+\infty} \mathscr{L} d\theta$$

then if we switch to λ:

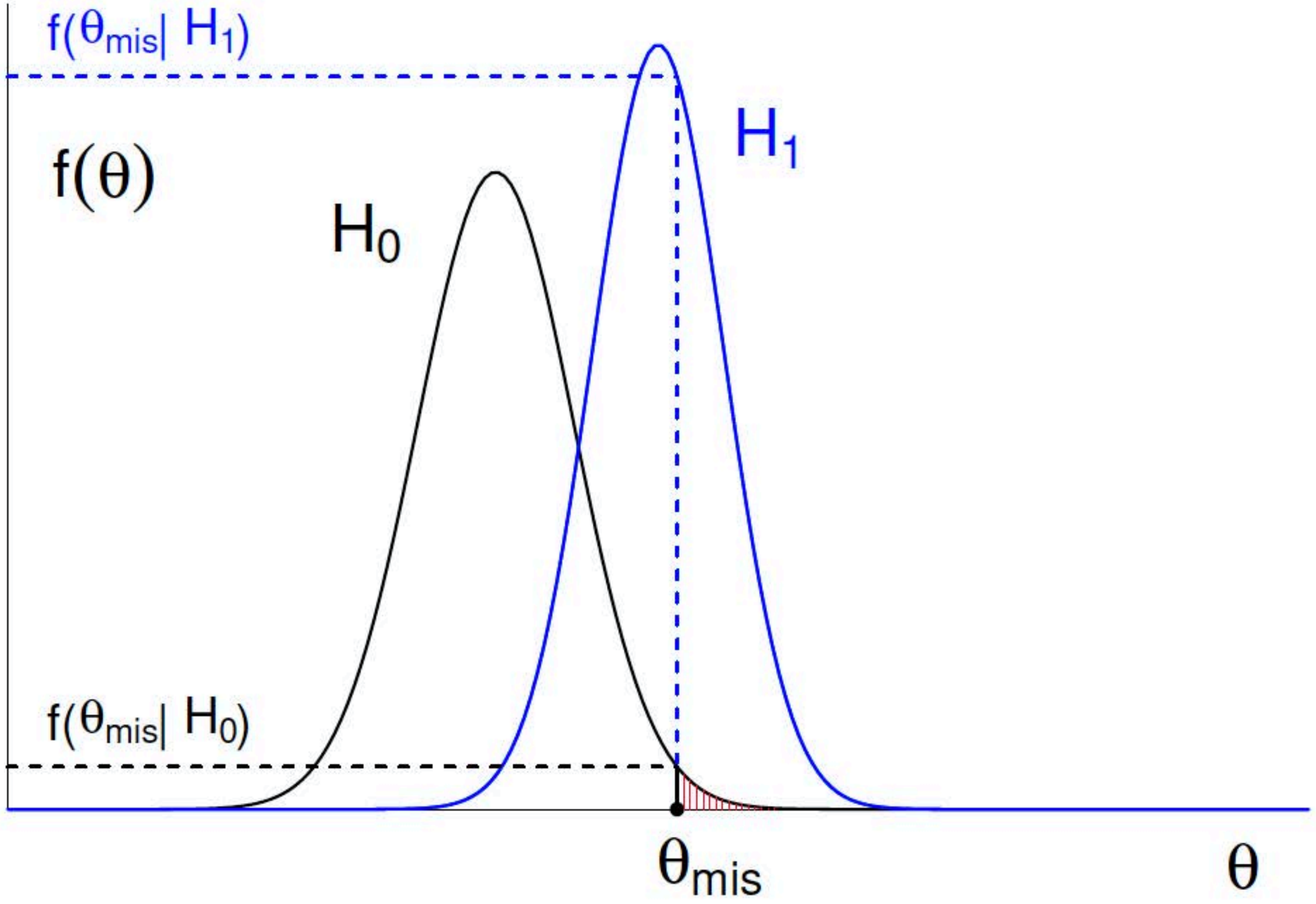
$$P(\lambda > \lambda') = \int_{\lambda'}^{\infty} \mathscr{L} d\theta \Big/ \int_{-\infty}^{+\infty} \mathscr{L} d\theta =$$

$$= \int_{\theta'}^{\infty} \mathscr{L} \frac{\partial \lambda}{\partial \theta} d\theta \Big/ \int_{-\infty}^{+\infty} \mathscr{L} d\theta \neq P(\theta > \theta'$$



p-values

- Why should the observation of θ_{mis} diminish our confidence on H_0?
- Because often we give some chance to a possible alternative hypothesis H_1



But if the alternative hypothesis H_1 is unconceivable, or hardly believable, the 'smallness' of the area is irrelevant

Confidence and significance

For historical reasons

- Confidence: CL = 1-α
- Significance: α.

Language of Hypothesis Testing:

Suppose that the pdf is known. Then if H_0 is true, the probability to get a measurement this far (or further!) is α.

Example:

Improvement among patients taking the treatment was significant at the 5% level' means that if the treatment does nothing, the probability of getting an effect this large (or larger) is 5% (or less).

Given a measurement, the corresponding **p-value** is the probability to get a larger value

→ The null hypothesis is rejected if the p-value is smaller than the significance

Bayesian point of view

- H0 = b (event is background)
- H1 = s (event is signal)

For each event test b. If b rejected, “accept” as candidate signal

- background efficiency $= \epsilon_b = P(x \in W | b) \equiv \alpha$
- signal efficiency = power $= \epsilon_s = P(x \in W | s) \equiv 1 - \beta$

To find purity of candidate signal sample, use Bayes’ theorem:

prior probability

signal region W

ε_s

$$P(\mathrm{s}|\mathbf{x} \in W) = \frac{P(\mathbf{x} \in W|\mathrm{s})P(\mathrm{s})}{P(\mathbf{x} \in W|\mathrm{s})P(\mathrm{s}) + P(\mathbf{x} \in W|\mathrm{b})P(\mathrm{b})}$$

posterior probability = signal purity

ε_b

Falsification

Largely considered the key to scientific progress (Popper)

$$\boxed{\textbf{if } C_i \nrightarrow E, \textbf{ then } E_{obs} \nrightarrow C_i}$$

Causes that cannot produce observed effects are ruled out ('falsified')

Analogous with method of the proof by contradiction of classical, deductive logic.

- Assume that a hypothesis is true
- Derive 'all' logical consequence
- If (at least) one of the consequences is known to be false, then the hypothesis is declared false

- What to do of all hypotheses that cannot be falsified?
- i.e. if nothing of what can be observed is incompatible with the hypothesis?

Example: H_i is a Gaussian $f(x \mid \mu_i, \sigma_i)$

- Given any pair or parameters $\{\mu_i, \sigma_i\}$, all values of x between $-\infty$ and $+\infty$ are possible
- Having observed any value of x, none of H_i can be, strictly speaking, falsified.

Falsification: the frequentist solution

- Way out: replace **impossible** with **improbable**!
- Mechanism deeply rooted in most people but not supported by logic
- Basically responsible of number of fake claims of discoveries
  - health
  - status of the planet

The falsification is now weakened

$$\boxed{\textbf{if}\ \ C_i \xrightarrow{small\ probability} E,\ \textbf{then}\ E_{obs} \xrightarrow{small\ probability} C_i}$$

➡ most likely false

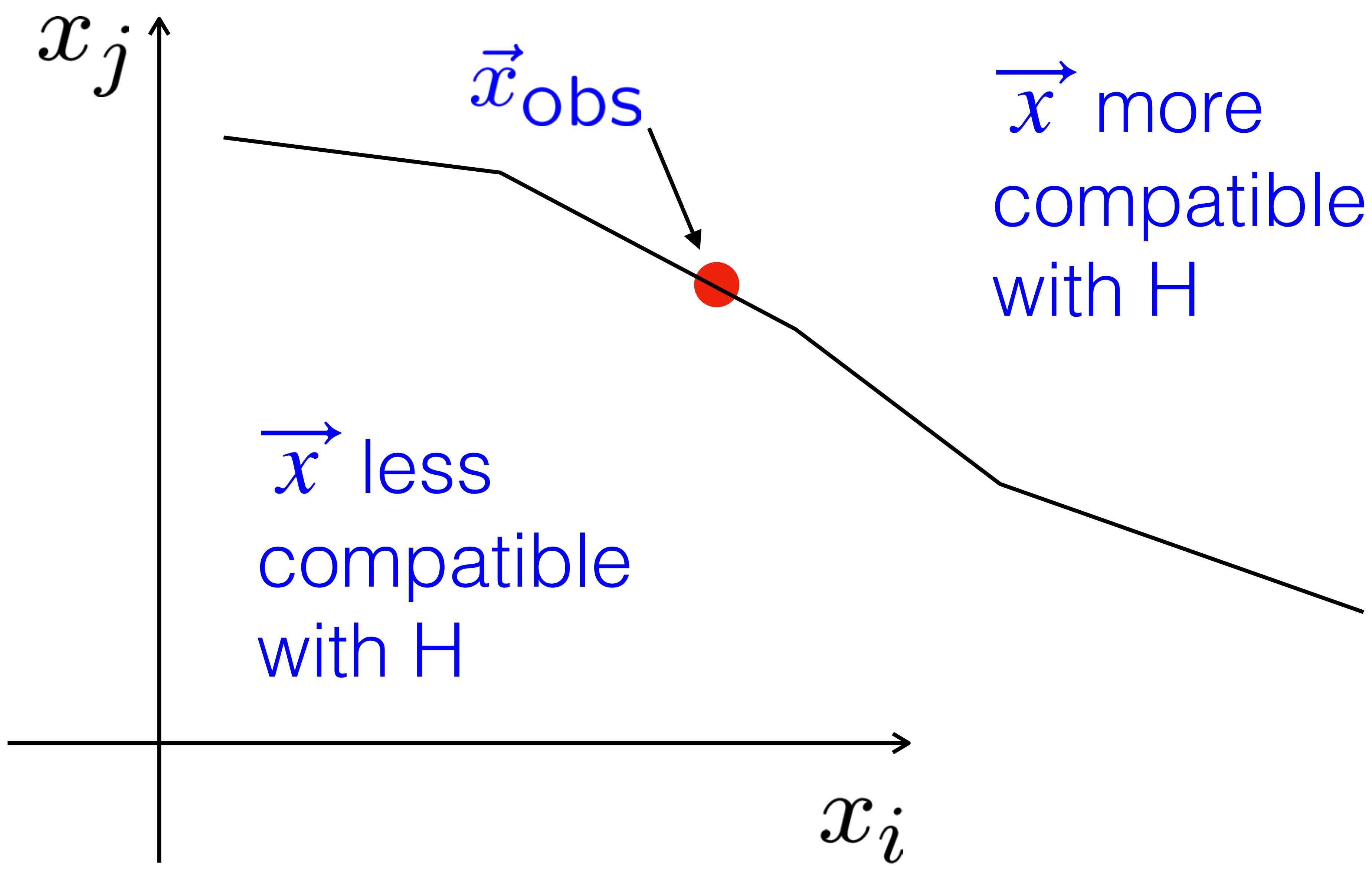
Testing significance / goodness-of-fit

Suppose hypothesis H predicts pdf $f(\overrightarrow{x}\,|\,H)$ for a set of observations $\overrightarrow{x} = \{x_1, \ldots, x_n\}$

If we observe a single point $\overrightarrow{x}_{obs}$ in this space, what can we say about the validity of H in light of the data?

Decide what part of the data space represents less compatibility with H than does the point $\overrightarrow{x}_{obs}$

This region therefore has greater compatibility with some alternative H′



x_j
$\vec{x}_{\mathrm{obs}}$
$\overrightarrow{x}$ more compatible with H
$\overrightarrow{x}$ less compatible with H
x_i

Test statistic and p-values

- Consider a parameter μ proportional to rate of signal process (μs+b).
- Define a function of the data (test statistic) q_μ that reflects the level of agreement between the data and the hypothesized value μ.
- Define q_μ so that higher values are increasingly incompatible with the data (more compatible with a relevant alternative).
- We can define critical region of test of μ by $q_\mu \geq$ const.

- Equivalently define the **p-value** of μ as:

$$p_\mu = \int_{q_{\mu,\mathrm{obs}}}^{\infty} f(q_\mu|\mu)\, dq_\mu$$

observed value of q_μ

pdf of q_μ assuming μ

Equivalent formulation of test: **reject *μ* if $p_\mu < \alpha$.**

Approximate confidence regions: LH function

Suppose we test parameter value(s) θ = (θ1, ..., θn) using the ratio

$$\lambda(\boldsymbol{\theta}) = \frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})} \qquad 0 \le \lambda(\boldsymbol{\theta}) \le 1$$

Lower λ(θ) means worse agreement between data and hypothesized θ.

Equivalently define

$$t_\theta = -2 \ln \lambda(\theta)$$

so higher t_θ means worse agreement between θ and the data.

Then p-value of θ:

$$p_{\boldsymbol{\theta}} = \int_{t_{\boldsymbol{\theta},\mathrm{obs}}}^{\infty} f(t_{\boldsymbol{\theta}}|\boldsymbol{\theta})\, dt_{\boldsymbol{\theta}}$$

need pdf

Confidence regions & Wilks' theorem

Wilks' theorem says (assuming proper conditions …)

$$f(t_\theta | \theta) \sim \chi^2_n$$

chi-square distribution for n d.o.f.
(n ≡ # of components in **θ** = (θ1, ..., θn))

Then p-value is

$$p_\theta = 1 - F_{\chi^2_n}(t_\theta)$$

where

$$F_{\chi^2_n}(t_\theta) \equiv \int_0^{t_\theta} f_{\chi^2_n}(t'_\theta)dt'_\theta$$

To find boundary of confidence region set pθ = α and solve for tθ:

$$t_\theta = F^{-1}_{\chi^2_n}(1-\alpha) = -2\ln\frac{L(\theta)}{L(\hat{\theta})}$$

$$\boxed{\begin{array}{l} \Delta^{1\sigma}_{n=1}\chi^2 = 1 \\ \Delta^{1\sigma}_{n=1}L = 1/2 \end{array}}$$

Confidence regions & Wilks' theorem (cont.)

Boundary of confidence region in θ space is where

$$\ln L(\theta) - \ln L(\hat{\theta}) = \frac{1}{2} F^{-1}_{\chi^2_n}(1-\alpha)$$

For example, for 1 – α = 68.3% and n = 1 parameter

$$F^{-1}\chi^2_1(0.683) = 1$$

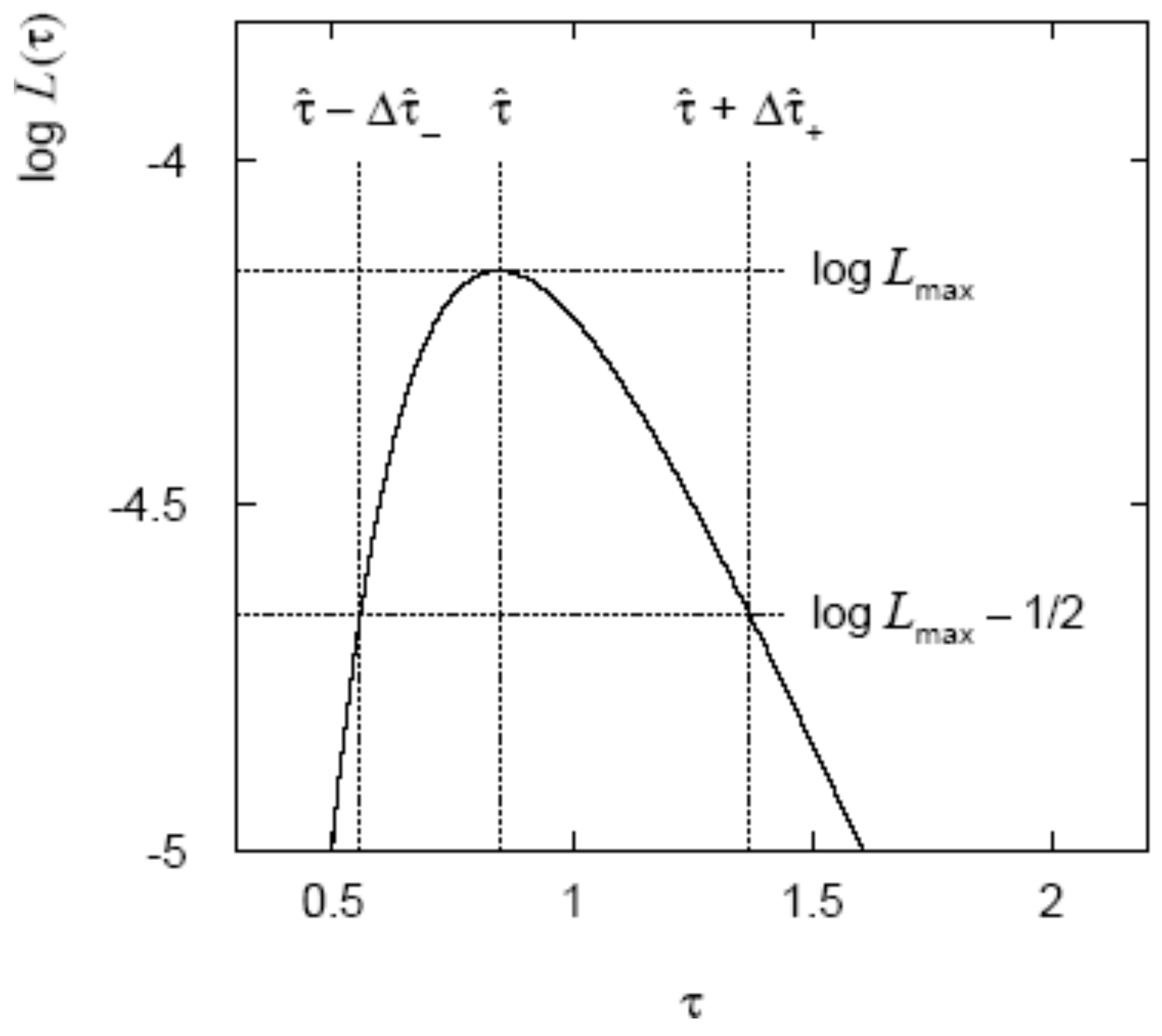
and so the 68.3% confidence level interval is determined by

$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{1}{2}$$

Same as recipe for finding the estimator's standard deviation, i.e. $[\hat{\theta} - \sigma_{\hat{\theta}}, \hat{\theta} + \sigma_{\hat{\theta}}]$ is a 68.3% CL confidence interval.

Exponential example with 5 events:

- n = 1 parameter,
- CL = 0.683



Parameter estimate and approximate 68.3% CL confidence interval: $\hat{\tau} = 0.85^{+0.52}_{-0.30}$

Nuisance parameters

Nuisance parameters

In general the models of the data, besides the parameter θ we are interested in, depends on a number of additional adjustable parameters ν: $p(x | \theta, \nu)$

Frequentist language: **nuisance parameters**

Bayes language: **systematic uncertainties**

Presence of nuisance parameter decreases sensitivity of analysis to the parameter of interest (e.g., increases variance of estimate)

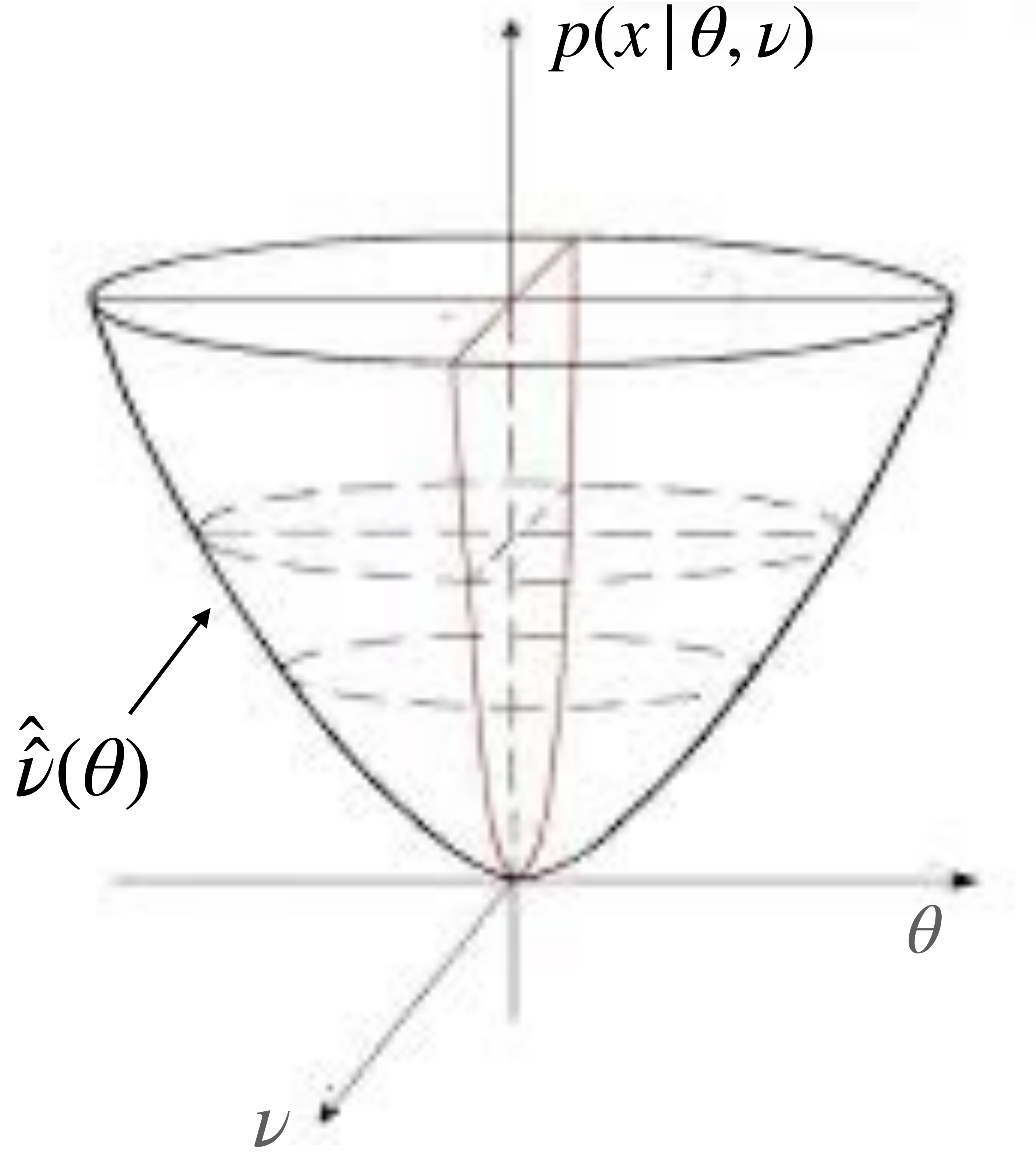
Profile likelihood ratio

Let's now consider a problem with likelihood L(θ, ν), where

$$\begin{aligned}\theta &= (\theta_1, \ldots, \theta_N) \\ \nu &= (\nu_1, \ldots, \nu_M)\end{aligned}$$

parameters of interest
nuisance parameters

Want to test point in θ-space



Define profile likelihood ratio:

"profiled" values of ***ν***

$$\lambda(\theta) = \frac{L(\theta, \hat{\hat{\nu}}(\theta))}{L(\hat{\theta}, \hat{\nu})} \quad \text{where} \quad \hat{\hat{\nu}}(\theta) = \underset{\nu}{\operatorname{argmax}} L(\theta, \nu)$$

and define $q_\theta = -2 \ln \lambda(\theta)$

Then Wilks' theorem says that distribution $f(q_\theta|\theta,\nu)$ approaches the chi-square pdf for N degrees of freedom for large sample (and regularity conditions), independent of the nuisance parameters ν

Example: counting experiment

Let's consider a spectrum $\mathbf{n}(E) = \{n_1 \ldots, n_N\}$ and assume the n$_i$ are Poisson distributed with expectation values $E[n_i] = \mu s_i + b_i$ parameter of interest

where $s_i = s_{tot} \int_{bin\ i} f_s(E;\theta_s)dE$ (signal) and $b_i = b_{tot} \int_{bin\ i} f_b(E;\theta_b)dE$ (background).

Suppose to have also some subsidiary measurement $\mathbf{m}(E) = \{m_1 \ldots, m_N\}$ such that $E[m_i] = u_i(\theta)$ (e.g. additional constraint on background).

Then the Likelihood function is $L(\mu, \boldsymbol{\theta}) = \prod_{j=1}^{N} \frac{(\mu s_j + b_j)^{n_j}}{n_j!} e^{-(\mu s_j + b_j)} \prod_{k=1}^{M} \frac{u_k^{m_k}}{m_k!} e^{-u_k}$

And the profile likelihood ratio

$$\lambda(\theta) = \frac{L(\theta, \hat{\hat{\nu}}(\theta))}{L(\hat{\theta}, \hat{\nu})}$$

defines the critical region of test of μ by the region of data space that gives the lowest values of λ(μ)

Important advantage of λ: its distribution becomes independent of nuisance parameters in large sample limit.

Markov Chain Monte Carlo (MCMC)

MCMC: the problem

- Imagine that we have a probability distribution for a set of parameters θ: p(θ|D, I)

D=data, I=model

- Then we need to marginalize over nuisance parameters ν:

$$p(\theta\,|\,D,I) = \int d\nu\, p(\theta,\nu\,|\,D,I)$$

- If the set {ν} is large this integral can become very expensive

We can integrate numerically using Monte Carlo sampling, but we waste time in regions of low probability

- Markov Chain Monte Carlo (MCMC) algorithms exploit a pure Bayesian approach
- Goal: draw samples from the PDF

$$p(\theta,\nu\,|\,D,I) = \frac{1}{Z} p(D\,|\,\theta,\nu,I)\, p(\theta,\nu\,|\,I)$$

where $Z = p(D|I)$ is the marginal evidence

- Since Z is independent of $\boldsymbol{\theta}$ and $\boldsymbol{\nu}$ we usually don't have to calculate it... which is good because it's expensive
- Once the samples produced by MCMC are available, the expectation value of a function of the model parameters $\boldsymbol{f}$ $(\boldsymbol{x})$ is

$$\langle f(\theta)\rangle = \int p(\theta\,|\,D,I) f(\theta) d\theta \approx \frac{1}{N}\sum_i^N f(\mathbf{x}_i)$$

- In MCMC, we randomly walk over positions $\boldsymbol{x}$ in the parameter space and draw samples $\boldsymbol{x}(t_i) = [\boldsymbol{\theta}_i, \boldsymbol{\nu}_i]$ from the distribution

Metropolis-Hastings algorhytm

- At each point in a Markov chain, **x**(t_i) depends only on the previous step **x**(t_{i-1}) according to the transition probability q(**x**(t + 1)|**x**(t)) (proposal distribution)

- The simplest MCMC algorithm is the Metropolis-Hastings method, which proceeds in two steps:
  1. Given **x**(t) sample a proposal position **y** from q(**y**|**x**(t))
  2. Accept this proposal with probability $\alpha = \min\left(1, \frac{p(y)}{p(x)}\frac{q(x\,|\,y)}{q(y\,|\,x)}\right)$

- In practice:
  1. Initialize x(0), set t = 0
  2. Sample y from q(y|x(t)) and u ~ Uniform(0,1)
  3. Evaluate r (the MH ratio in α); if u≤r then x(t+1)→y; otherwise,x(t+1)→x(t)

Reversible Markov Chains

- Let’s consider the Markov chain $(x_t)_{t \geq 0}$ with the transition kernel (y|x) and the stationarity condition

$$\pi(y) = \int p(y \mid x)\pi(x)dx$$

- The chain is reversible (i.e. it satisfies the balance condition) if it exists a distribution π(x) s.t.

$$p(x \mid y)\pi(y) = p(y \mid x)\pi(x)$$

- Then π(x) is stationary

$$\int p(y \mid x)\pi(x)dx = \int p(x \mid y)\pi(y)dx = \pi(y)\int p(x \mid y)dx = \pi(y)$$

since $\int p(x \mid y)dx = 1$

Balance condition and reversibility

- Now we can demonstrate that the balance condition is satisfied (the chain is reversible and admits a stationary distribution).
- To this end we have to write the transition kernel p(y|x) which will depend on q(y|x) and the MH criterion (p(x|y) ➝ α(y|x)q(y|x))
- If y≠x

$$
\begin{aligned}
p(y\,|\,x)p(x) &= \alpha(y\,|\,x)\; q(y\,|\,x)\; p(x) \\
&= \min\left(1, \frac{p(y)}{p(x)}\frac{q(x\,|\,y)}{q(y\,|\,x)}\right) q(y\,|\,x)\; p(x) \\
&= \min\left(q(y\,|\,x)\; p(x), q(x\,|\,y)\; p(y)\right) \\
&= \min\left(\frac{q(y\,|\,x)}{q(x\,|\,y)}\frac{p(x)}{p(y)}, 1\right) q(x\,|\,y)\; p(y) = \alpha(x\,|\,y)\; q(x\,|\,y)\; p(y) \\
&= p(x\,|\,y)\; p(y)
\end{aligned}
$$

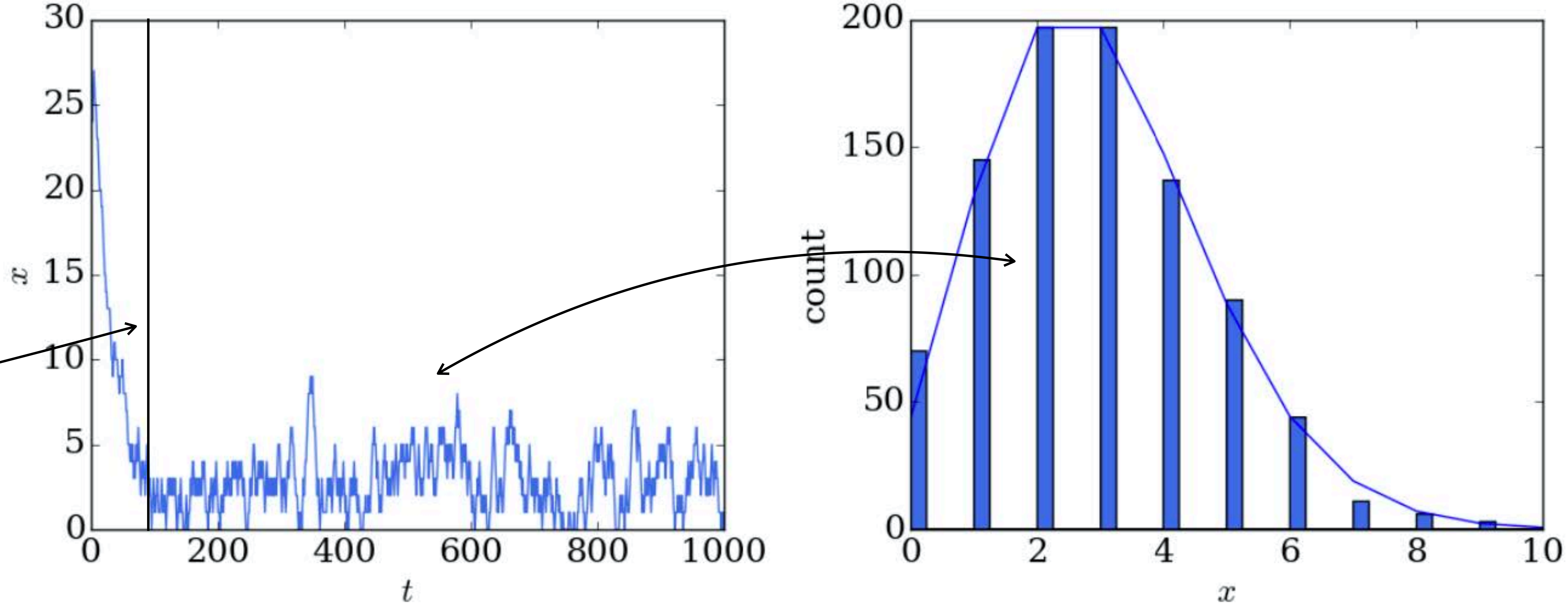
Example: sampling from Poisson distribution

- We want to sample from the 1D PDF $p(x|D,I) = e^{-\lambda}\lambda^{x}/x!$.
- Let's choose $q(y|x_t)$ to be a simple random walk defined by the uniform distribution

1. Given x_t, pick a random number $u_1 \sim$ Uniform(0, 1)
   - if $u1 > 0.5$: propose $y = x_t + 1$
   - otherwise, $y = x_t - 1$
2. Compute the ratio $r = p(y|D,I)/p(x_t|D,I) = \lambda^{y-x}x!/y!$
3. Generate a second random number $u_2 \sim$ Uniform(0, 1)
   - If $u_2 \leq r$ accept $x_{t+1} = y$
   - otherwise $x_{t+1} = x$

proposal distribution $q(y|x)$

- (left) sequence of the first 1000 samples $\{x_t\}$
- (rightt) histogram of x_t for $t > 100$
- **"burn in"**: first 100 samples (to be discarded)



Autocorrelation function (ACF)

- Autocorrelation tells you how much each step in the time series depends on the value of previous steps

- If the transition probability is independent of t then the ACF should fluctuate around zero. This is what happens after the burn-in
- During the burn-in, the ACF is roughly exponential in shape

$$\rho_{xx}(h) \sim \exp\left(-\frac{h}{\tau}\right)$$

where τ is called the time constant

- Larger τ means that the MCMC takes longer to converge, so the goal is to choose a proposal distribution that minimizes τ
- Empirically, you can estimate τ from the data, and start using the data when t is several multiples of τ
- For our Poisson example, τ ≈ 23 samples, so to be safe we've started using the data at t = 4τ ≈ 100

Tips for shopping:

- When implementing a calculation, it is always better to use logarithms rather than actual values to avoid hitting numeric limits. If the actual PDF is needed we exponentiate at the end of the calculation.
- Better using the definition n! = Γ(n + 1) instead of using Stirling's approximation $\ln n! \approx n \ln n - n$

MCMC efficiency

The Metropolis-Hastings algorithm works because it reaches an equilibrium state after the burn-in.
In particular, the transition probabilities obey the detailed balance equation, which characterizes a Markov Chain

- A number of issues have to be decided when running an MCMC:
  - What is the length of the burn-in period? i.e., when can we start trusting the data?
  - When do we stop the Markov Chain? i.e., how do we know if we've sufficiently sampled the parameter space?
  - How do we choose a suitable proposal distribution that gives a reasonable acceptance rate for transitions xt → xt+1?
- There is a large literature about optimizing Markov Chain Monte Carlo
- An MCMC that takes forever to burn-in or which accepts few transitions isn't worth much
- Current state of the art: affine-invariant samplers

- There are various tricks to speed up MCMC and ensure that it explores as much of the parameter space as possible:
  - One common approach is to define multiple chains (or "walkers") that have different starting points and proceed independently
  - If the sampled PDF is very peaked or multimodal, this might still not be enough to push explore all parts of the parameter space

Another example: sampling from joint posterior

- Let's consider a joint distribution p(x_1,x_2|D,I) in two parameters x_1 and x_2.

- Assume a 2D Gaussian for the pdf (double peaked structure):

$$p(x_1, x_2 \,|\, D, I) = \frac{1}{2}[G(\mu_1, \boldsymbol{\Sigma}_1) + G(\mu_2, \boldsymbol{\Sigma}_2)]$$

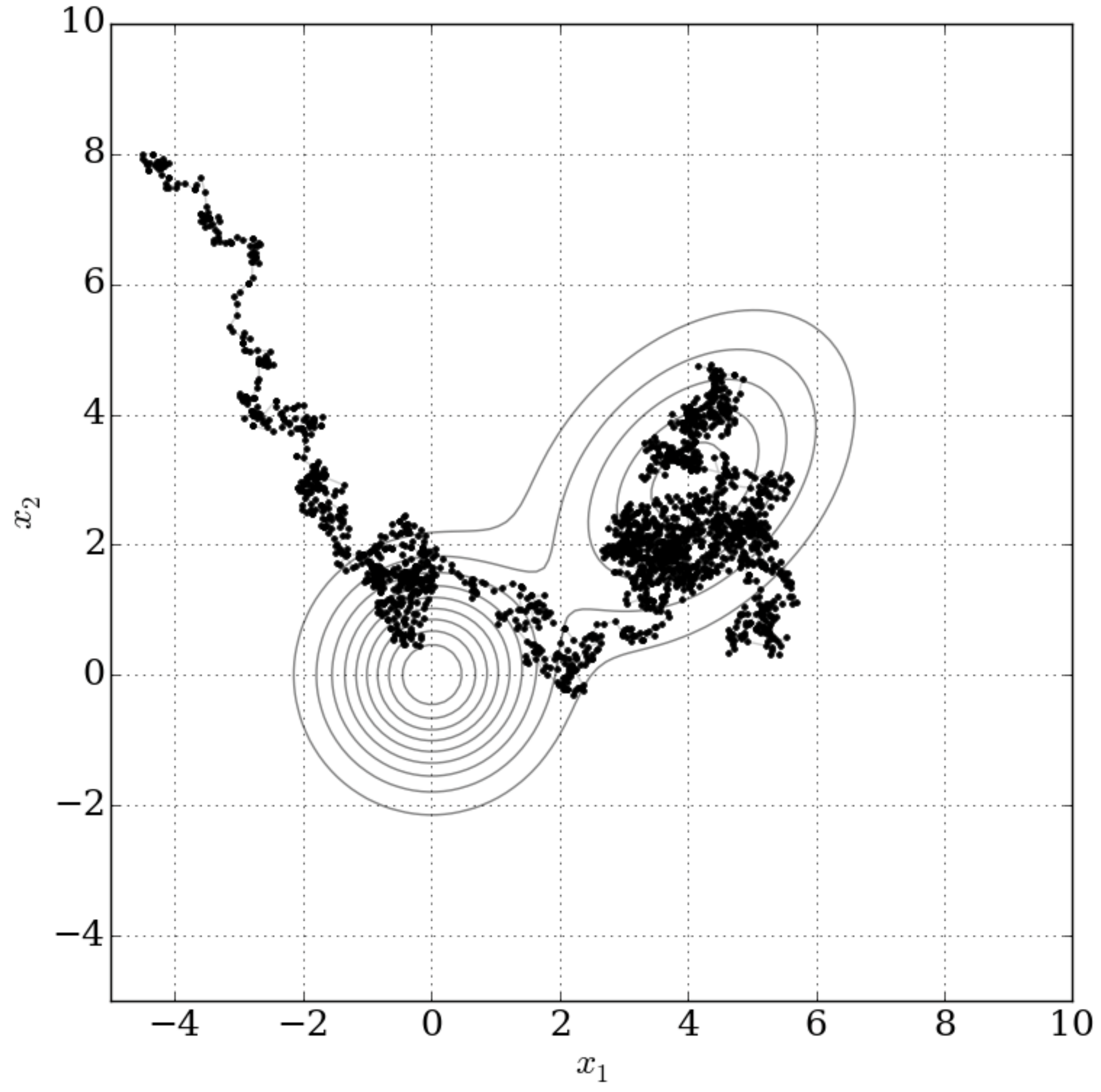
with μ_1=(0,0), μ_2=(4,3), and $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 0.8 \\ 0.8 & 2 \end{pmatrix}$

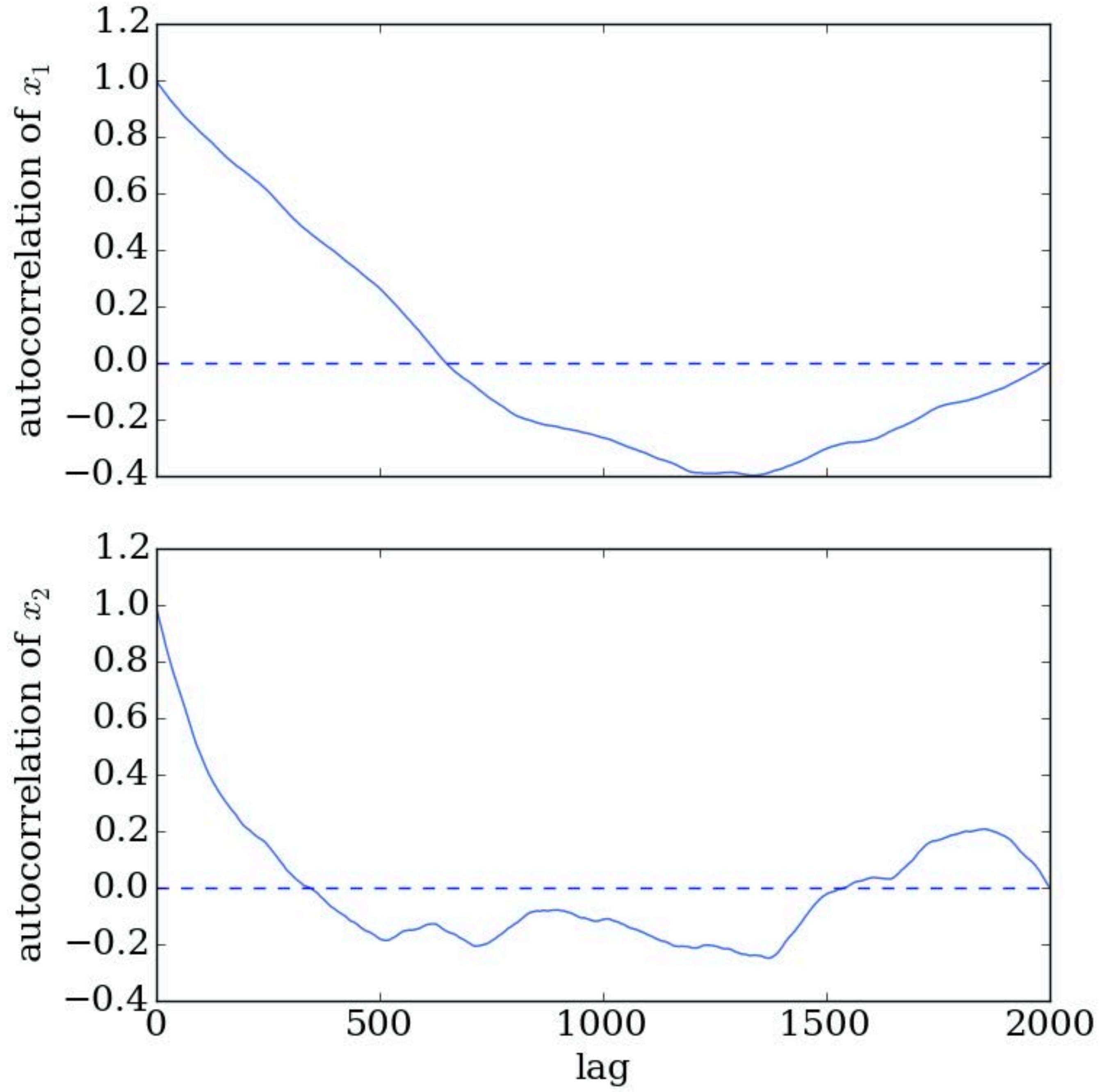
- Use a unimodal 2D Gaussian for the proposal density function (proposal distribution)

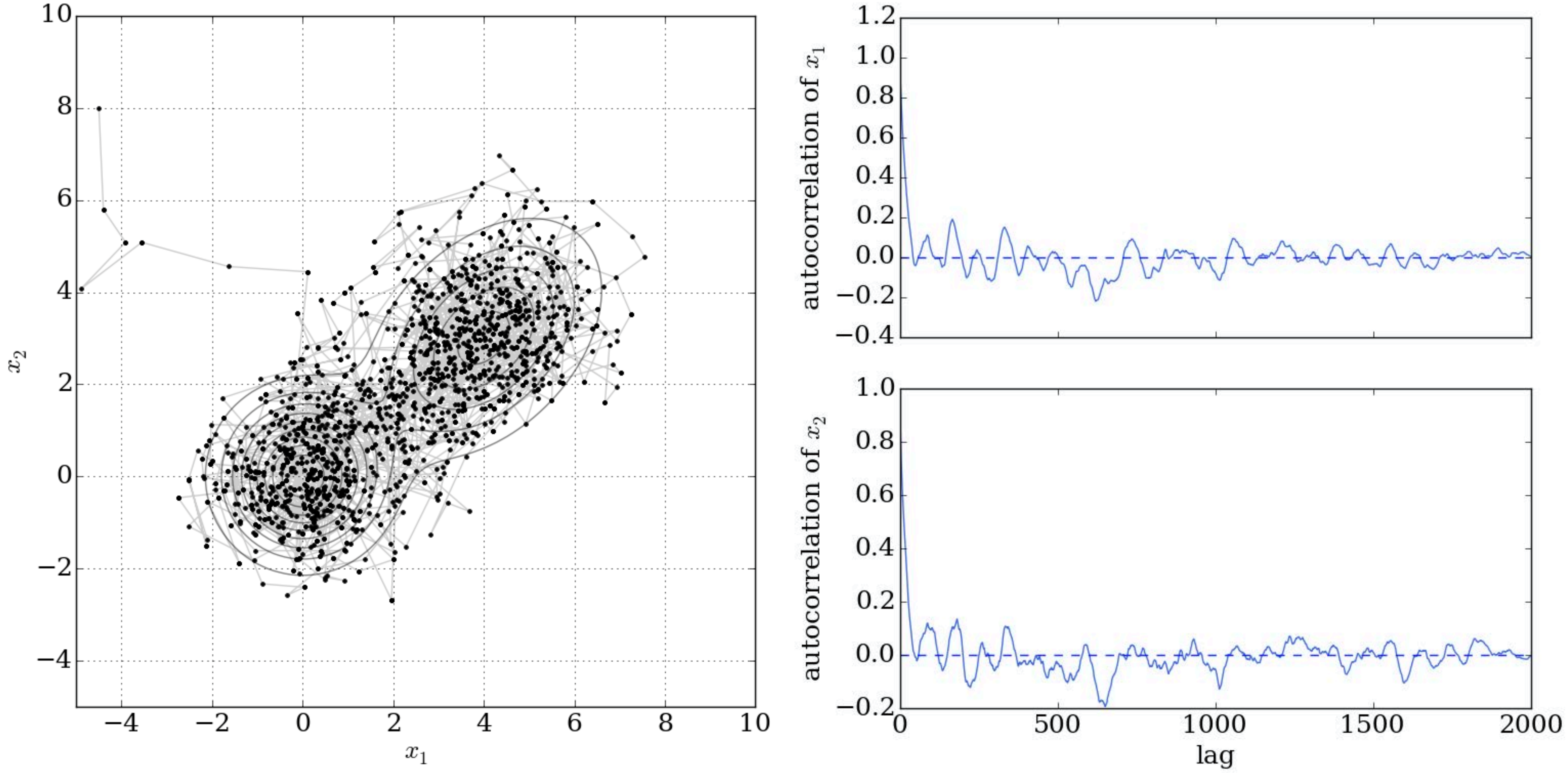
$$q(\mathbf{y} \,|\, \mathbf{x}) = G(\mu = \mathbf{x}, \boldsymbol{\Sigma}_q),$$

$$\boldsymbol{\Sigma}_q = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$$

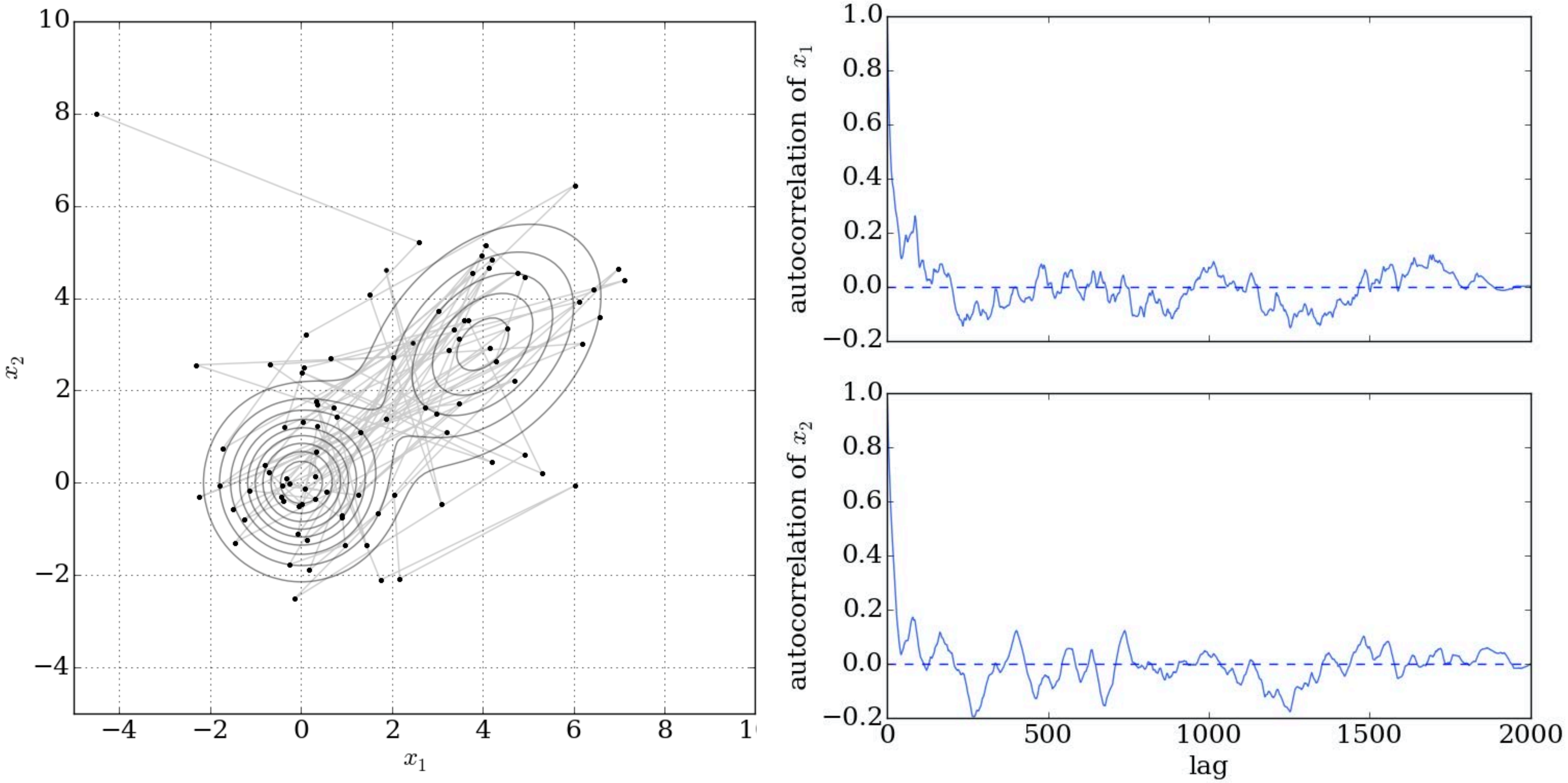
- $\mathbf{x}_0$ = (-4.5,8)
- σ = 0.1
- Very long autocorrelation time
- Acceptance probability ~ 95%





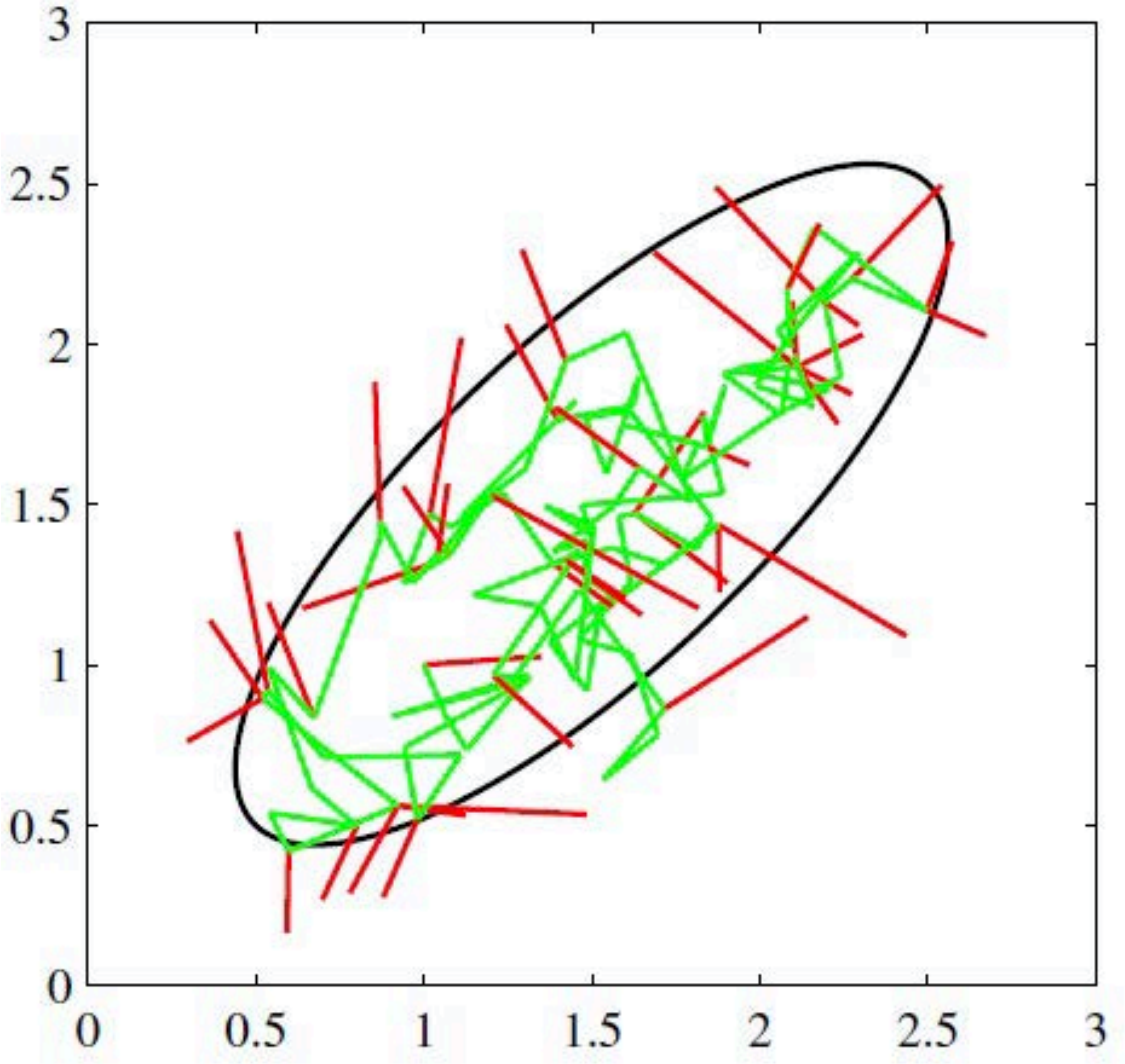


- $\mathbf{x}_0$ = (-4.5,8)
- σ = 1
- Faster convergence
- Acceptance probability ~ 60%

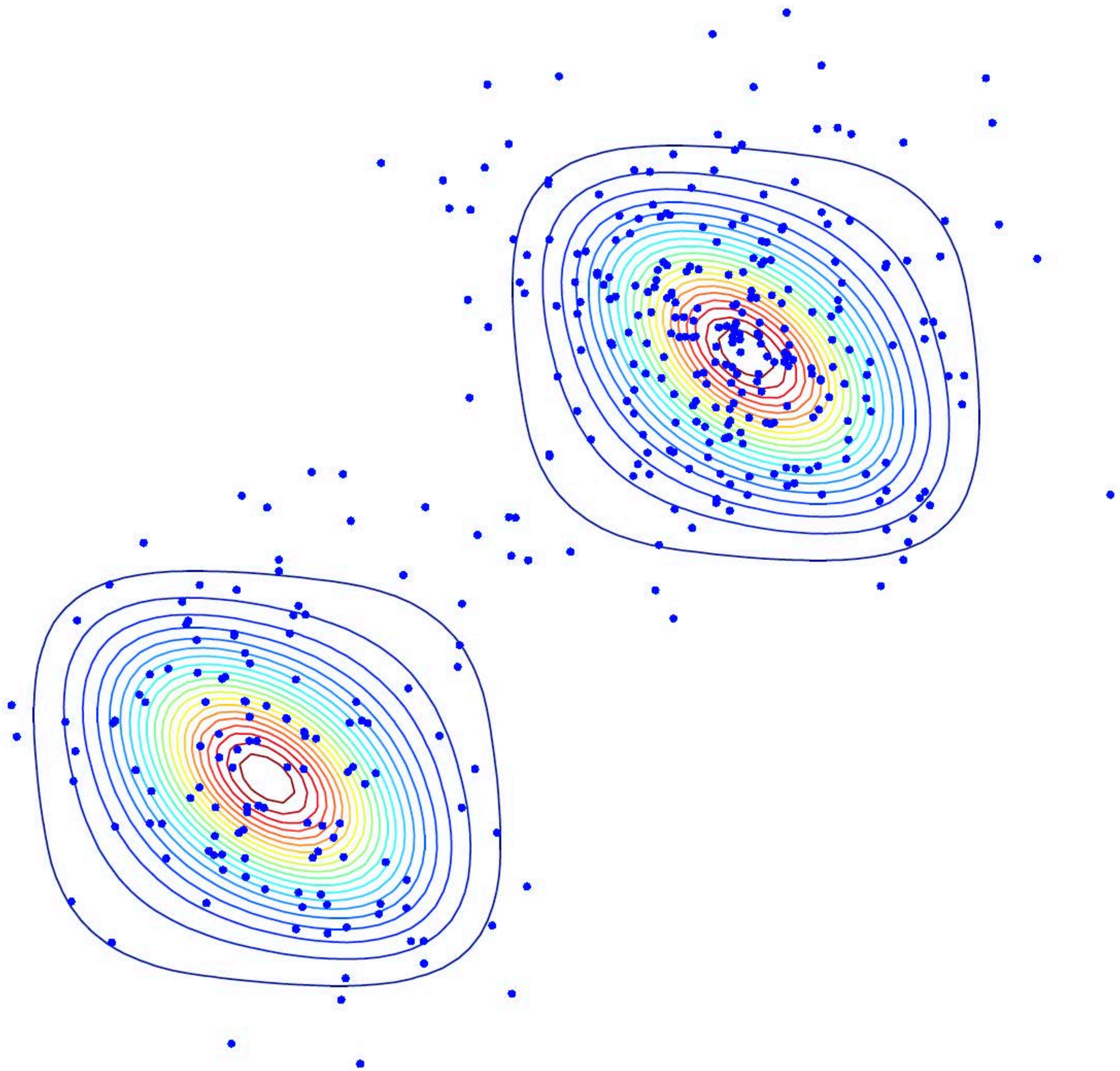


- $\mathbf{x}_0$ = (-4.5,8)
- σ = 10
- Fast convergence
- Acceptance probability ~ 5%

MCMC paths



A simple illustration using Metropolis algorithm to sample from a Gaussian distribution whose one standard-deviation contour is shown by the ellipse. The proposal distribution is an isotropic Gaussian distribution whose standard deviation is 0:2. Steps that are accepted are shown as green lines, and rejected steps are shown in red. A total of 150 candidate samples are generated, of which 43 are rejected (from Bishop's *Pattern Recognition and Machine Learning*).



Metropolis-Hastings samples from a bi-variate distribution p(x1; x2) using a proposal ~q(x0jx) = N (x0 x; I).
We also plot the iso-probability contours of p. Althoug p(x) is multi-modal, the dimensionality is low enough and the modes sfficiently close such that a simple Gaussian proposal distribution is able to bridge the two modes. In higher dimensions, such multi-modality is more problematic (from Barber’s *Bayesian reasoning and machine learning*)

MCMC: Gibbs sampling

Let's consider: $p(x) = p(x_1, \ldots, x_n), \qquad x = (x_1, \ldots, x_n)$
assumption: we can sample from 1-DIM conditionals (priors)

$$p(x_i | x_{\neg i}) \;\; x_{\neg i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$$

- Pick $k \in \{1,\ldots,n\}$ (round-robin, uniform random, …)
- Set $x_j^{(t+1)} = x_j^{(t)}$ for j≠k
- Sample $x_k^{(t+1)} \sim p(x_k | x_{\neg k}^{(t)})$

  with $q_k(y|x) = p(y_k | x_{\neg k})$ when $y_{\neg k} = x_{\neg k}$ and 0 otherwise

no jumps allowed between different indexes

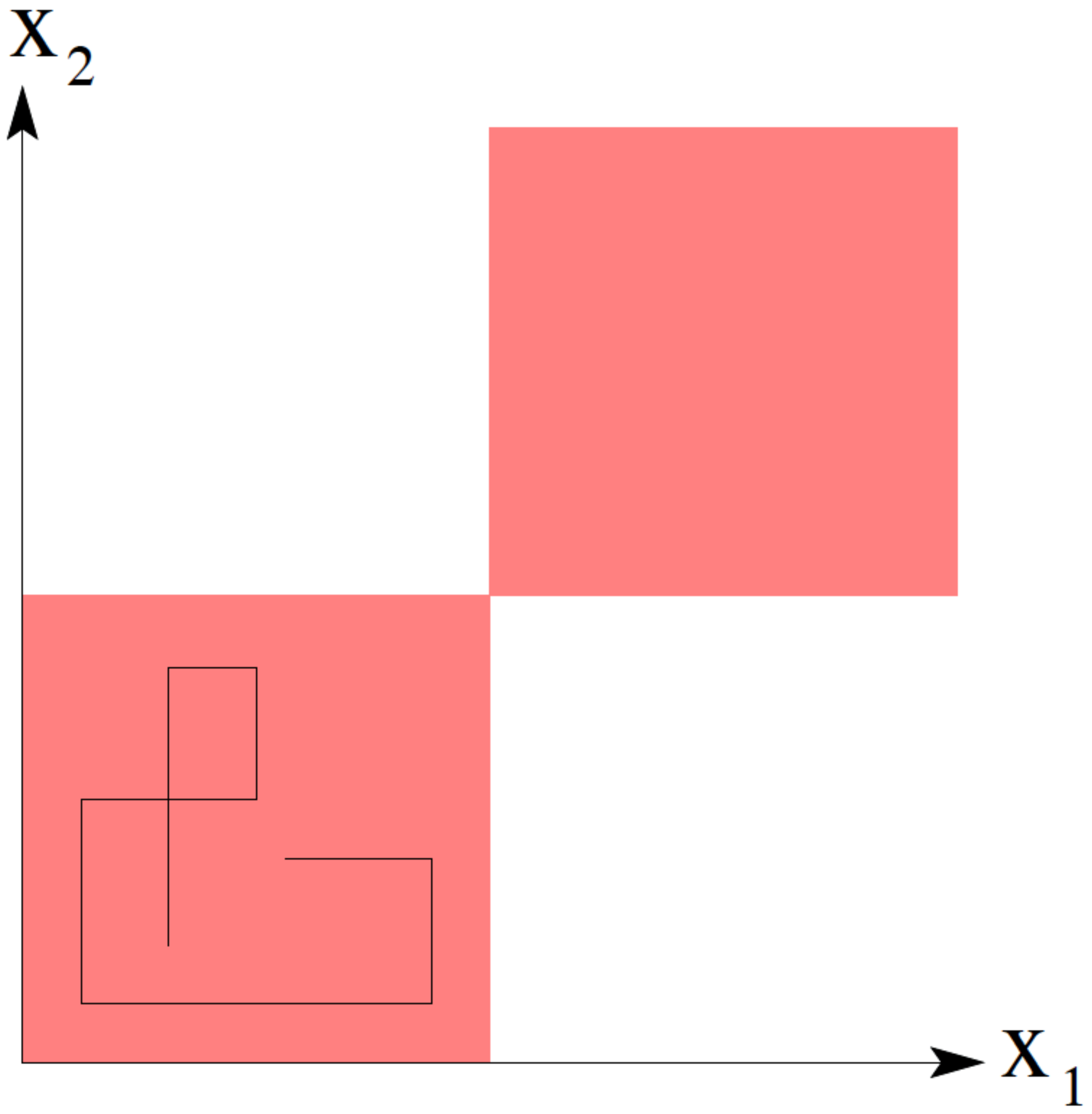
This implies $\alpha_k(y|x) = 1$ for the acceptance probability:

$$\alpha_k^{(MH)} = \frac{p(y)q_k(x|y)}{p(x)q_k(y|x)} = \frac{\cancel{p(y_k|y_{\neg k})}\cancel{p(y_{\neg k})}}{\cancel{p(x_k|x_{\neg k})}\cancel{p(x_{\neg k})}} \cdot \frac{\cancel{p(x_k|y_{\neg k})})}{\cancel{p(y_k|x_{\neg k})}} = 1$$

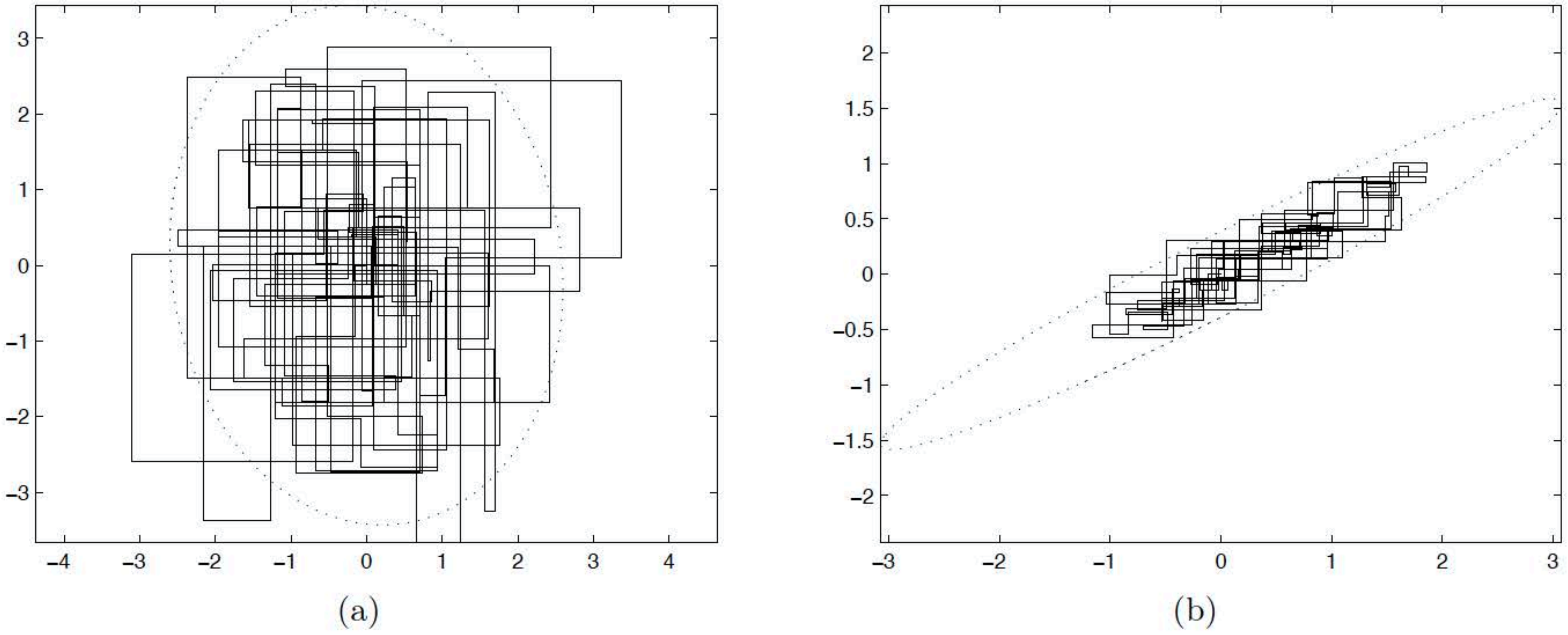
probability decomposition of p(y_k) and p(x_k)

- Expensive
- A number of known issues but often is the only choice:
  - ergodicity (connect all points)
  - slow convergence in case of correlations

Gibbs issues



A two dimensional distribution for which Gibbs sampling fails. The distribution has mass only in the shaded quadrants. Gibbs sampling proceeds from the lth sample state (x_1^l, x_2^l) and then sampling from $p(x_2 \,|\, x_1^l)$, which we write (x_1^{l+1}, x_2^{l+1}) where $x_1^{l+1} = x_1^l$. One then continues with a sample from $p(x_1 \,|\, x_2 = x_2^{l+1})$, etc. If we start in the lower left quadrant and proceed this way, the upper right region is never explored



(a)

(b)

Two hundred Gibbs samples for a two dimensional Gaussian. At each stage only a single component is updated. (a): For a Gaussian with low correlation, Gibbs sampling can move through the likely regions effectively. (b): For a strongly correlated Gaussian, Gibbs sampling is less effective and does not rapidly explore the likely regions

from Barber’s *Bayesian reasoning and machine learning*

“Local minima”

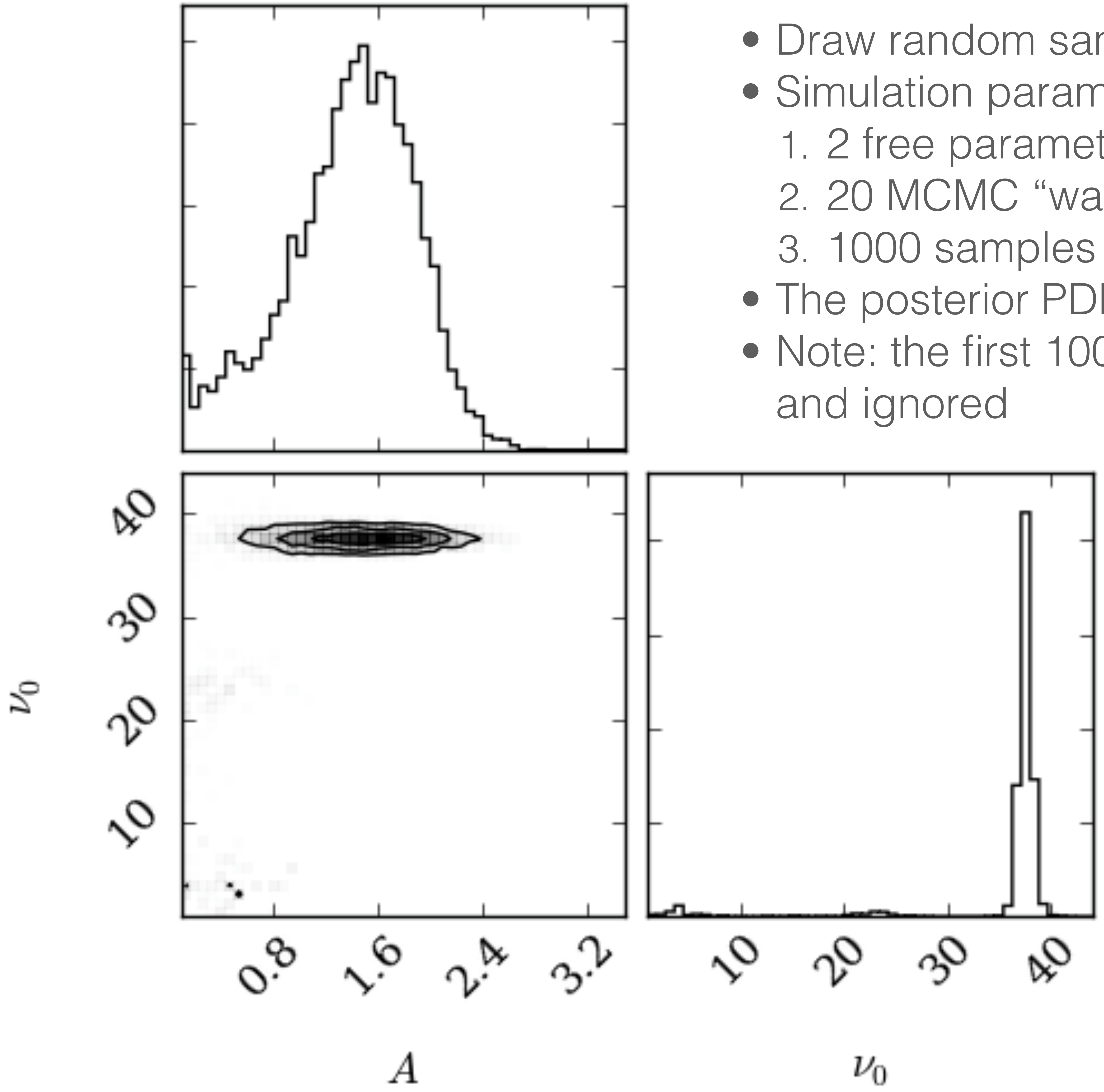
- The situation when MCMC become stuck in a local mode is similar to the situation in parameter estimation when a minimizer gets stuck in a local minimum
- **Solution:**
  - Create a series of progressively “flatter” distributions using a temperature parameter T (or β = 1/T)
  - As T → ∞ and β → 0, the distribution will flatten and more of the paramter space can be explored
  - Given a posterior $p(\mathbf{x}\,|\,D, I) \propto p(\mathbf{x}\,|\,I)p(D\,|\,\mathbf{x}, I)$ we can construct a flattened distribution using β ∊ [0, 1]: $\pi(\mathbf{x}\,|\,D, \beta, I) = p(\mathbf{x}\,|\,I)p(D\,|\,\mathbf{x}, I)^{\beta} = p(\mathbf{x}\,|\,I)\exp(\beta\ln[p(D\,|\,\mathbf{x}, I)])$
- With π(x|D,β,I) we can use a set of discrete values β = {1,β$_2$,...,β$_m$} in parallel

- **Parallel Tempering:**
  - multiple copies of the MCMC are run in parallel, each with a different temperature βi
  - As the simulations run, pairs of adjacent simulations on the temperature ladder are allowed to swap their parameter states with probability $r = \min\left(1, \dfrac{\pi(\mathbf{x}_{t,i+1}\,|\,D, \beta_i, I)\pi(\mathbf{x}_{t,i}\,|\,D, \beta_{i+1}, I)}{\pi(\mathbf{x}_{t,i}\,|\,D, \beta_i, I)\pi(\mathbf{x}_{t,i+1}\,|\,D, \beta_{i+1}, I)}\right)$

Algorithm:

1. Propose a swap every ns iterations, and proceed with the swap if u1 ~ Uniform(0,1) ≤ 1/ns
2. Randomly pick simulation i to swap its state with simulation i+1
3. Accept the swap if u2 ~ Uniform(0,1) ≤ r
4.

A final example with tampering



- Draw random samples from $p(A,\nu_0|D,I)$
- Simulation parameters:
  1. 2 free parameters: A, ν_0 with independent σ's
  2. 20 MCMC "walkers"
  3. 1000 samples
- The posterior PDF is shown at left with the marginal distributions of A and ν_0
- Note: the first 100 samples from each walker were treated as burn-in data and ignored

CUORE: background model with JAGS

- Reconstruction of all possible sources of background that affect CUORE data
- Identify the combinations of source volume, isotope and contaminant distribution that can impact our background
- Model the contribution of each possible contaminant with MC simulations
- Combine all contributions with a Bayesian fit
- Can include a priori information on contaminants
- No maximization algorithm (numerical approach - MCMC)
- Easily includes all uncertainties and correlations

Pro	Con
Marginalised distributions	SLOW
Correlations	
Systematic errors (nuisance parameters)	
Priors	

Bayes: $\tilde{p}(\theta\,|\,\mathrm{data}) = \mathrm{p}(\mathrm{data}\,|\,\theta)\,\mathrm{p}(\theta)$

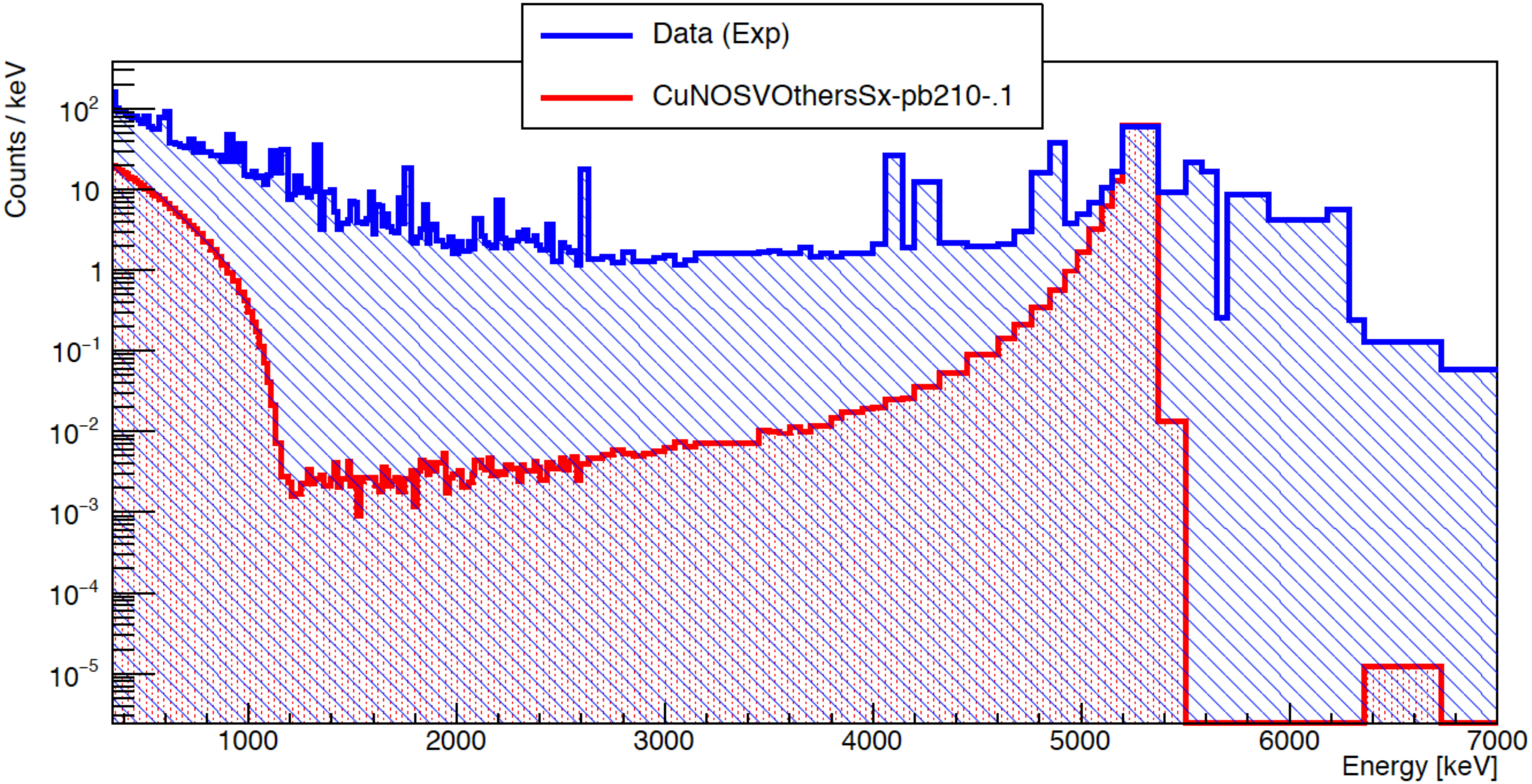
- data is the experimental spectrum while **θ** is normalisation vector of the MC simulated spectra
- priors (when non flat) come from previous measurements
- contaminations must be > 0 and their contribution cannot exceed the observed spectrum
- n_i is the number of experimental counts in bin i
- $m_{i,M}$ is the number of MC counts in bin i from source M
- Likelihood: $\prod_i P(n_i\,|\,\theta) = \frac{e^{-\lambda_i}\lambda_i^{n_i}}{n_i!}$, where $\lambda_i = \sum_M \lambda_{i,M} = \theta_M \sum_M m_{i,M}$

Gibbs sampling: JAGS

- JAGS (Just Another Gibbs Sampler) is Just Another Bayesian Tool which can be used to sample a multivariate likelihood
- Allows the definition and the solution of our problem
- Not ROOT-compatible natively, it operates with text files only (→ BAT)
- Binned fits to improve speed
- Data preparation: need to prepare data and MC histograms with the required binning in JAGS-compatible text files.
- Need to inform JAGS about statistical model and MCMC parameters
- Finally extracts useful numbers from the JAGS output (correlated posterior pdf) and creates plots
-

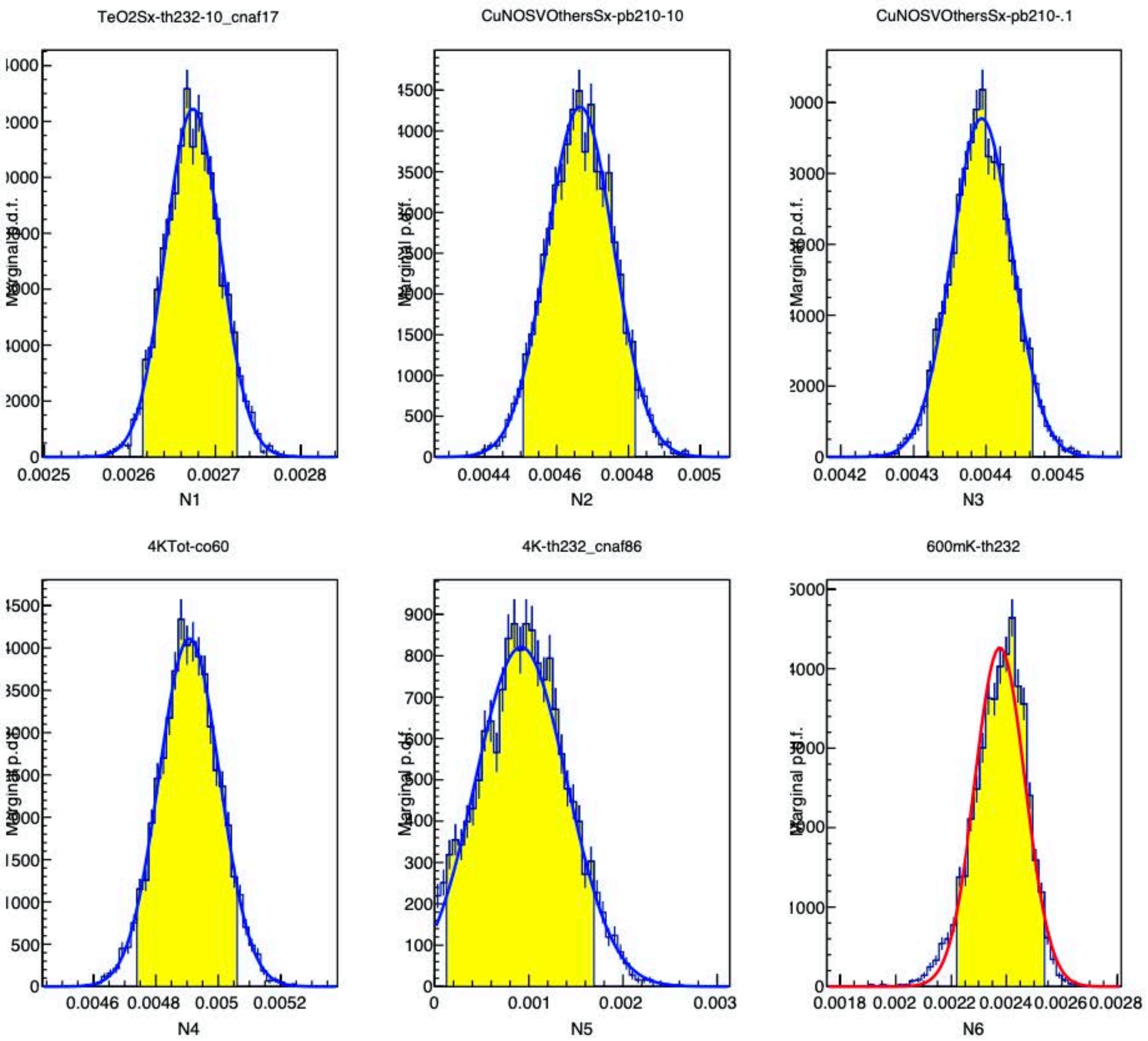
Priors (when no measurement is available)

1. Find the maximum MC normalization factor allowed by the data
2. Multiply it by a safe factor
3. Uniform (flat) prior between 0 and this value
4.



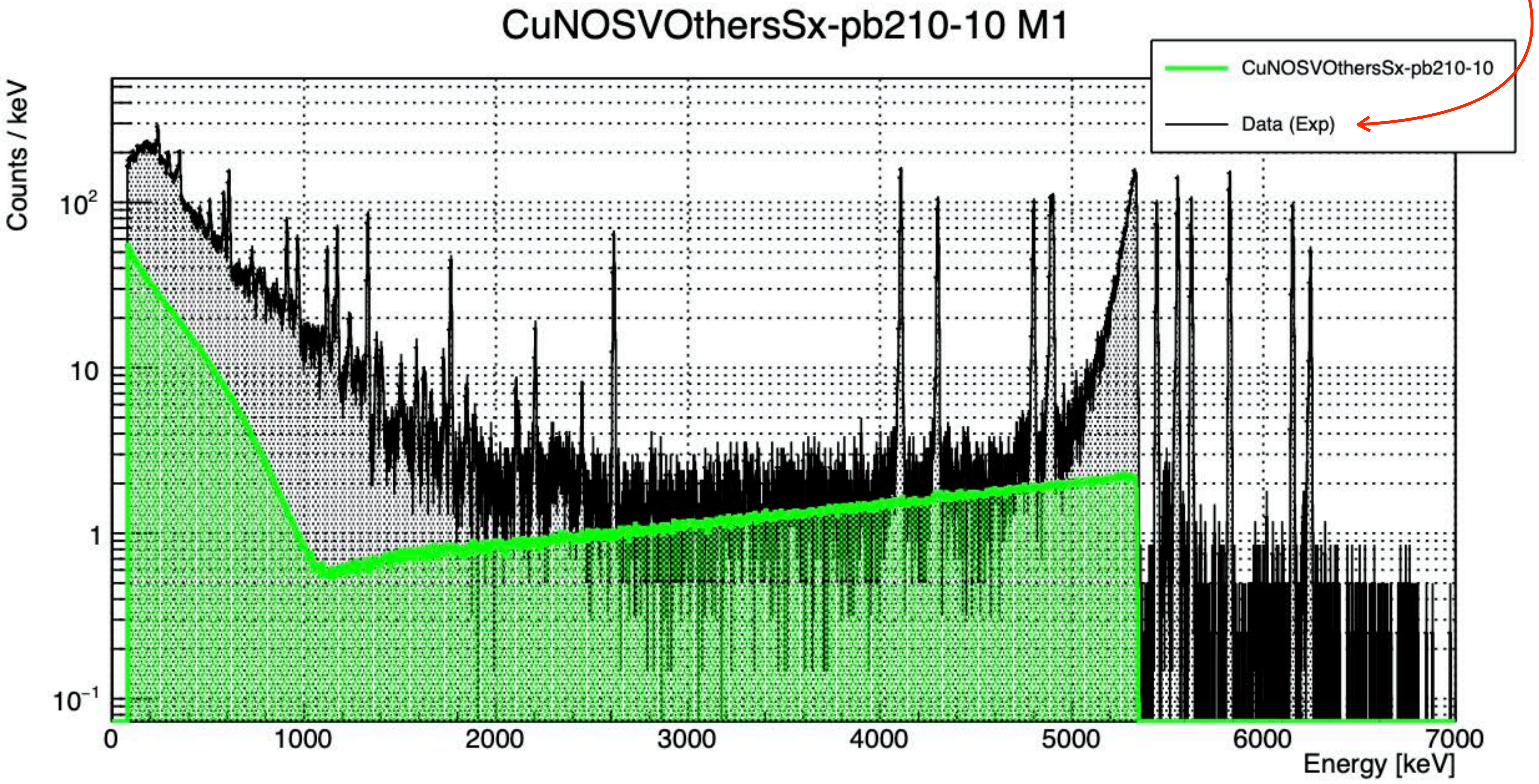
Results (1/3)

Marginalized distributions



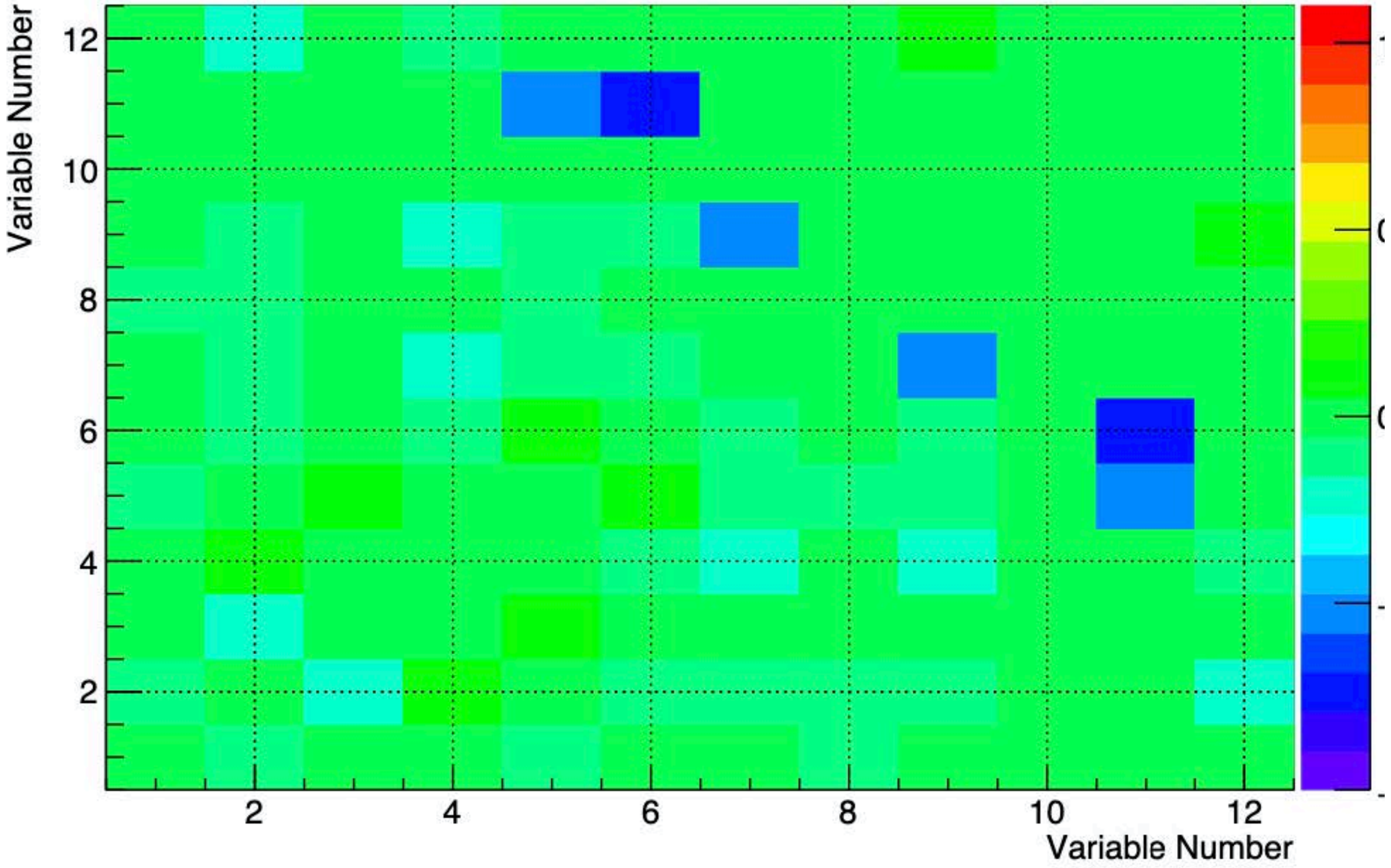
Normalized spectra

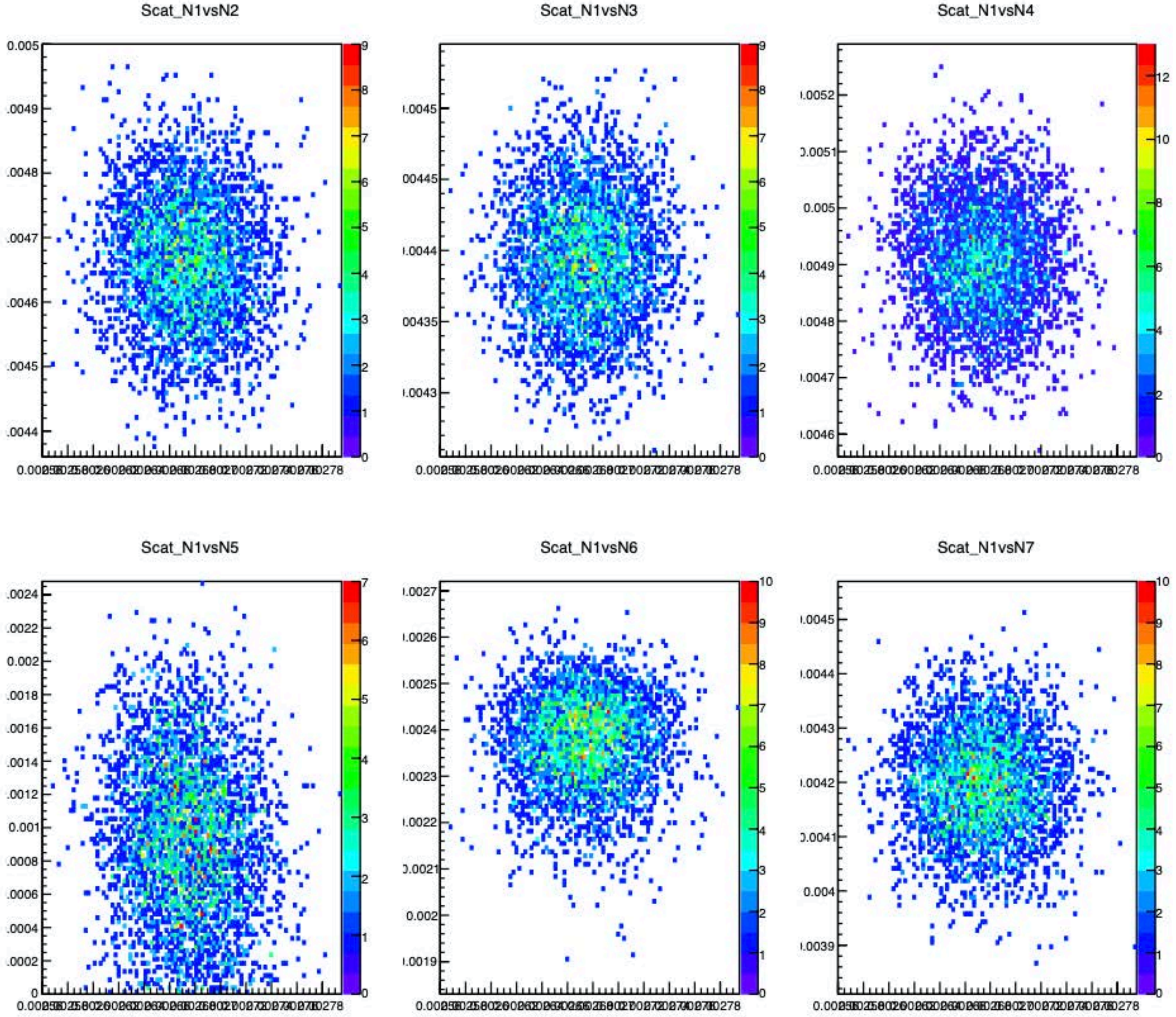
Exp in reality is MC for this test exercise



Results (2/3)

Correlation plots

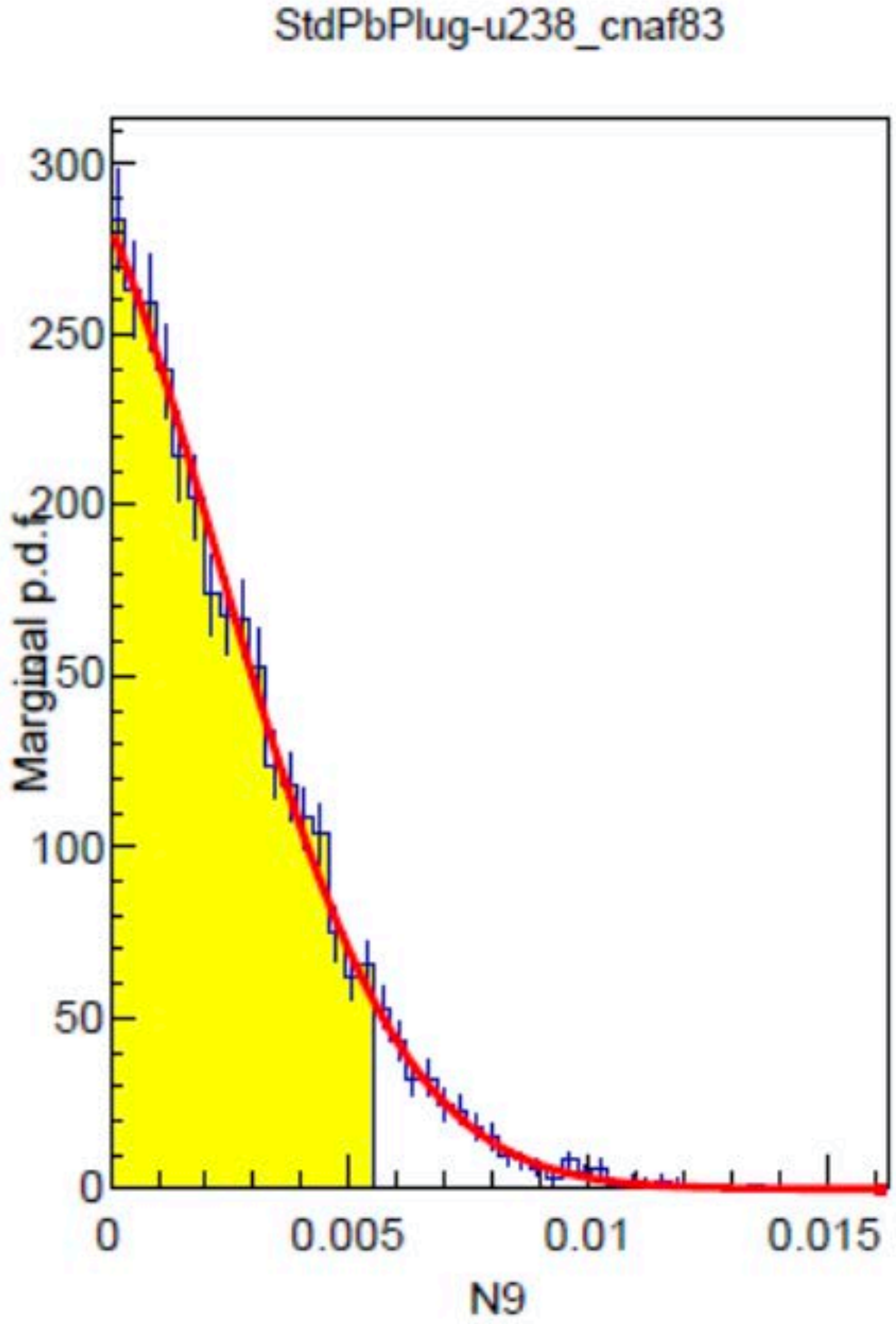




Results (3/3)

From marginalised distributions:

- Credibility intervals: from gaussian parameters
- Limits (1-α C.I. q(α)): $\int_0^{q^{\alpha)}} p(\theta_k)d\theta_k$



Systematic uncertainties

Systematic errors

"Systematic errors arise from neglected effects such as incorrectly calibrated equipment."

- NOOOO: FALSE!
- A neglected effect is a MISTAKE! **A MISTAKE is not an ERROR**
- So what are systematic errors?

Systematic Error: reproducible inaccuracy introduced by faulty equipment, calibration, or technique

Bevington

Systematic effects is a general category which includes effects such as background, scanning efficiency, energy resolution, angle resolution, variation of couner efficiency with beam position and energy, dead time, etc. The uncertainty in the estimation of such as systematic effect is called a systematic error

Orear

- Analysis of your results involves a whole set of numerical factors: efficiencies, magnetic fields, dimensions, calibrations...
- Occasionally these are implicit: these are especially dangerous
- All these numbers have an associated uncertainty.
- **These uncertainties are the systematic errors.**
- They obey all the usual error laws, but they affect all measurement

Systematic errors: examples

- The magnetic field in **p = 0.3 B R**
  - Calorimeter energy calibration 'Jet energy scale'
  - Detector efficiency
  - ...
- Effect of uncertainty in B on the error matrix for two momentum measurements:

$$V = \begin{bmatrix} 0.3^2 B^2 \sigma_1^2 + 0.3^2 R_1^2 \sigma_B^2 & 0.3^2 R_1 R_2 \sigma_B^2 \\ 0.3^2 R_1 R_2 \sigma_B^2 & 0.3^2 B^2 \sigma_2^2 + 0.3^2 R_2^2 \sigma_B^2 \end{bmatrix}$$

- Errors on p_1 and p_2 as given by simple combination of errors.
- Also covariance/correlation term.
- Errors in B affect both momenta measurements the same way

Most common:

- Instrument(s) parameters:
  - Zero setting
  - Linearity
  - Response
- Constants
- …

- If you can't think of them for your experiment, ask a colleague …
- … possibly with a talent for criticism. (not difficult to find one …)

However … not an easy task!

- Many properties of the reconstruction don't work through simple algebra.

Standard (?) procedure:

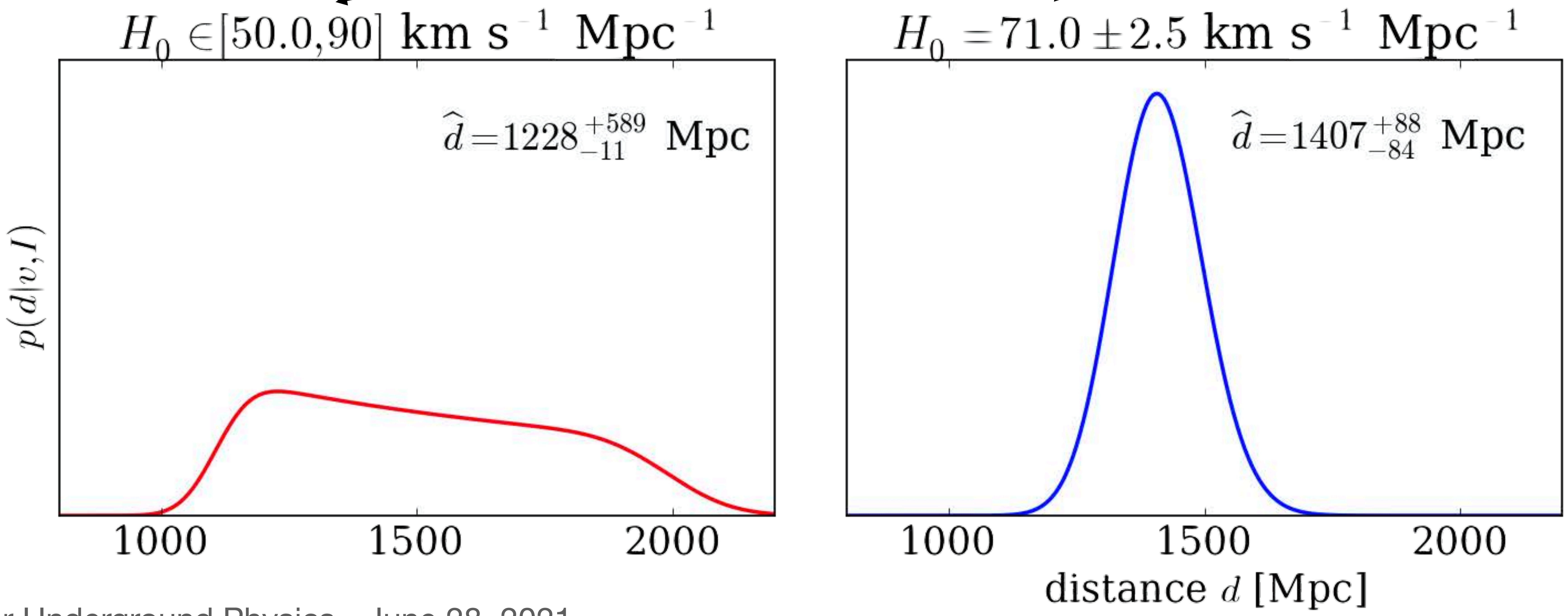
- Example: background to your signal simulated by Monte Carlo containing several (?) adjustable parameters...
- Work numerically. Run standard MC, then adjust parameter by +σ and repeat, -σ and repeat. Read off error from shift in result
- If you can convince yourself that the 3 points are a straight line then do so

Systematic errors: Bayesian viewpoint

In the Bayesian framework, since uncertainties reflect the degree of belief rather than just the spread of repeated measurements, it’s straightforward to incorporate a “parameter uncertainty” through its prior.

Example: compute the distance to a galaxy given its recession velocity v taking into account uncertainties in H_0.

➡ Bayes solution: select a prior on H_0 (Uniform, Gaussian,)



Systematic uncertainty and bias

- The bias is the difference between the “true” and measured value
- “Bias is equivalent to systematic uncertainty”
  ➡ True when we know there is a bias but its exact size is unknown.

- **However there are other possibilities:**
  - Bias is known, with known size; so we correct for it → not a systematic (known knowns)
  - Bias is known, but exact size is unknown → systematic uncertainty (known unknowns)
  - Bias is unknown and unsuspected → nothing to be done (unknown unknowns)

- If you are unaware of a systematic effect in your data, you can get internally consistent results with an impressive goodness-of-fit and still be completely wrong
- Unfortunately no magic recipe …

- Usual suggestions:
  - Split the data into subsets and analyze them separately
  - Vary cuts, bin sizes, etc. and explore the effect on the results
  - Change parameterizations or fit techniques
  - Perform independent analyses and check differences in outcomes

Systematic errors: some advice

- Check your result by altering features which should make no (significant) difference. This adds to its credibility
- Run on subsets of the data (time etc)
- Change cuts on quality and kinematic quantities
- Check that a full blown analysis on simulated data returns the physics you put in
- Repeat until you (and your colleagues or review committee) really believe
- Challenge data using Known Inputs (create a simulated dataset with known inputs and see if the inputs are recovered

- If repeating with some difference in technique gives a different result …
  - You have to decide whether this is significant.
  - “Within Errors” may be overgenerous since results share the same data (or some of it)
  - Subtraction in quadrature is one way:
    - Basic result: 12.3 ± 0.4. Check: 11.7 ± 0.5 Compare difference 0.6 against $\sqrt{0.5^2 - 0.4^2} = 0.3$

- If the analysis passes the check with a small difference
  - Tick the box and move on
  - Do not fold that small difference into the systematic error
- If the analysis fails the check
  - Check the check
  - Check the analysis and find the problem
  - Maybe convince yourself that this 'harmless' change could cause a systematic shift and devise an appropriate error
  - Do not fold the difference into the systematic error

A very last remark: there is no correct procedure for incorporating a check that fails, but folding it into the systematics is probably wrong and should be avoided unless there is no alternative

Systematic uncertainties: conclusions

- Systematic uncertainties are a frequentist concept; for a Bayesian, there is no distinction and all such uncertainties can be dealt with using marginalization
- Still, it's useful to break out uncertainties into statistical and systematic components, as this (usually) makes clear which part of the error bar depends on how much data we took
- When conducting an experiment, one tries to identify systematic effects before, during, and after data-taking.
- There is no recipe for doing this right but there are some "best practices" that good researchers try to follow
- After all efforts have been made to eliminate systematic effects, the remaining uncertainties become systematic uncertainties.
- It is important not to inflate systematics, but in the real world, sometimes you do have to cut your losses and go with a reasonable uncertainty

The 6 Barlow commandments

Thou shalt never say ‘systematic error’ when thou meanest ‘systematic effect’ or ‘systematic mistake’

Thou shalt not add uncertainties on uncertainties in quadrature. If they are larger than chickenfeed, get more Monte Carlo data

Thou shalt know at all times whether thou art performing a check for a mistake or an evaluation of an uncertainty

Thou shalt not not incorporate successful check results into thy total systematic error and make thereby a shield behind which to hide thy dodgy result

Thou shalt not incorporate failed check results unless thou art truly at thy wits’ end

Thou shalt say what thou doest, and thou shalt be able to justify it out of thine own mouth, not the mouth of thy supervisor, nor thy colleague who did the analysis last time, nor thy mate down the pub.

Do these, and thou shalt prosper, and thine analysis likewise

Roger Barlow